Cover Sheet for MA Dissertation

Deadline:	Date submitted:
14 September 2016	12 September 2016

Surname/Family Name:	Other name(s):	Degree and name of programme:
Vargova	Maria	MA in Language Testing (by distance)

Course No. / Title:	Supervisor's name:	Word Count:
MALTD Dissertation	Dr. Luke Harding	19 878
Dissertation Title:		
Developing partnership between teaching and testing listening: exploring the degree of alignment between teaching and testing listening comprehension in the context of NATO STANAG 6001 exam preparatory courses		

I declare that this submission is my own work. I have not submitted it in substantially the same form towards the award of a degree or other qualification. It has not been written or composed by any other person and all sources have been appropriately referenced or acknowledged. In the case of electronically submitted work, I consent to this work being stored electronically and copied for assessment purposes, including the department's use of plagiarism detection systems in order to check the integrity of assessed work.

Signed: 

Please tick as appropriate:
☑ I am willing for this (anonymised) piece of work to become part of an Essay Bank
☐ I am not willing for this (anonymised) piece of work to become part of an Essay Bank

Supervisor's Comments:

Additional Points to Note:
☐ More proof-reading was required
☐ Make sure you follow the method of referencing outlined on the General Info website
☐ Ensure that you number your pages
☐ Ensure that you use 1.5 or double line spacing

Mark:

Penalty:

Final Mark:

Marker's Initials:

IMPORTANT NOTES
1. In exceptional circumstances (e.g. illness), it is possible to apply for an extension to coursework deadlines. Without an extension, ten marks will be removed from the agreed final mark for up to three working days late.
2. A copy of this document will be kept in your file in the Department.
3. All marks are provisional until confirmed by the appropriate Board of Examiners.

LANCASTER UNIVERSITY

Department of Linguistics and English Language

DEVELOPING PARTNERSHIP BETWEEN TEACHING AND TESTING LISTENING

exploring the degree of alignment between teaching and testing listening comprehension in the context of NATO STANAG 6001 exam preparation courses

MARIA VARGOVA

Dissertation submitted in partial fulfilment of the requirements for the

M.A. in Language Testing (by distance)

September 2016

19 878 words

Abstract

Since Slovakia became a full member of NATO in 2004, having a required level of proficiency in English has become essential for future career prospects of military as well as civilian personnel working for the Armed Forces of the Slovak Republic. While test-takers are generally successful in attaining Level 1 (survival level) from speaking, reading and writing, their average success rate on a Level 1 listening test is relatively low. As a meaningful link between what is tested and what is taught is believed to affect students' test performance (e.g., Hughes, 2003; Popham, 2001), this study attempted to explore the degree of alignment between the listening construct tested by a listening section of Level 1 exam designed in accordance with the NATO Standardization Agreement 6001 and the listening construct taught in the courses aimed at preparing their participants for taking the exam.

To explore the research issue, a mixed methods convergent parallel design consisting of a pre-test and a post-test version of a questionnaire, a listening test, teacher interviews and classroom observations was employed. Altogether, 51 students attending Level 1 preparatory courses and 8 teachers teaching in the courses participated in the study. The data gathered were analyzed both quantitatively and qualitatively. The comparison between the construct tested and the construct taught was based on Buck's (2001) framework for describing listening ability because that was the framework used to define the construct targeted by the test.

Although several weaknesses were found in the way listening comprehension was taught and tested at Level 1, the findings indicated that there was a clear alignment between the listening sub-skills and strategies taught in Level 1 courses and those measured by Level 1 listening test. Based on the findings, implications and suggestions for educational as well as testing practice were suggested.

Acknowledgements

I would first like to recognise my supervisor, Dr. Luke Harding, for guiding me throughout the whole process of the study. I am grateful for all his insightful suggestions, comments and tips, which helped me greatly in shaping the direction of the study.

Next, I would like to thank all the other members of the Department of Linguistics and English Language in Lancaster for putting together such a great master programme as MA in Language Testing (by distance) undoubtedly is.

I would also like to express my sincere appreciation to the management of the Language Institute and my two senior colleagues, Janka Balíková and Danka Bruteničová, for providing me with all the documents and statistical data I needed to carry out this study.

Thanks are also extended to all the study participants as the study would not have been possible without their taking time to share their views and experiences.

I am also grateful to all my teacher and tester colleagues for their great support, but most of all to Gabika Repová and Sonka Gajdošová for being an abundant source of positive energy and inspiration.

A sincere appreciation belongs to my close family and friends who supported me throughout my whole studies. My very special thanks, however, goes to my dearest sister Zuzka for her incredible willingness to review each of my module assignments as well as my thesis. I greatly appreciate her kindness and help.

Finally, my deepest thanks of all goes to my beloved husband Maroš and our two children, Linda and Eduard, for their endless patience, encouragement and love.

Table of contents

1 Introduction

> Without a semblance of alignment, nothing hangs together.
>
> (Baker, 2004, p. 5)

Three years after the fall of a communist regime in Czechoslovakia, the Slovak Republic became an independent state. Since its establishment in January 1993, accession to the North Atlantic Treaty Organization (NATO) became one of the Slovakia's main foreign and security policy priorities. In 1999, following the fall of Mečiar's government, criticized for violating democratic norms, Slovakia's chances of NATO membership improved, a clear indication of which was NATO's approval of a Membership Action Plan (MAP) for Slovakia. In addition, a newly formed governing coalition ratified a Program for Preparation for NATO Membership, which was based on and fully reflected the conditions of accession of Slovakia to NATO laid down by MAP. The document defined the tasks Slovakia had to undertake to meet NATO's accession requirements. Improving language proficiency of military personnel working for the Armed Forces of the Slovak Republic (AF SR) was listed as one of the main defence and military priorities (Bebler, 2010).

Although official languages at NATO are both French and English, English is recognized by many as NATO's operational language (Dubeau, 2006; Solak, 2012). That is the main reason why the prime focus of foreign language training provided to members of the AF SR has been on teaching and testing English. The body to which the role of helping to build and develop language proficiency and fluency of military and civilian personnel working for the AF SR has been assigned is the Language Institute (LI), working under the General Staff of the AF SR. Besides ensuring language training for military personnel, one of the main tasks of the LI is also to develop language exams in accordance with the NATO Standardization Agreement 6001 Edition 5 (STANAG 6001), which is an agreement requiring NATO full and aspiring member states to measure, record and report language ability of their military personnel using STANAG 6001 scale. (www.nato.int)

Passing the exam based on STANAG 6001 (STANAG 6001 exam) has important implications for the career development of military as well as civilian members of the AF SR, including promotion, deployment abroad or acquisition of a job within NATO structures. This has especially been true since 2004, when Slovakia formally became a full member of NATO. As a great majority of Slovak Armed Forces members realize that the difference between having and not having the needed STANAG 6001 language proficiency level is the difference between having and not having good career prospects, they try to achieve the level by attending STANAG 6001 exam preparation courses, provided by the LI with an intention to prepare the participants of these courses for taking STANAG 6001 exam.

Being a tester involved in the development of STANAG 6001 exam and having experience with teaching STANAG 6001 preparation courses, what inspired me to lay the focus of my MA dissertation study on looking into foreign language listening comprehension in the context of STANAG 6001 courses at Level 1 (L1 courses) was that a listening section of STANAG 6001 exam at Level 1 (L1 listening test) has long been the most difficult part of the exam for the participants of L1 courses. While, according to the statistics of the test results obtained on STANAG 6001 exam at Level 1 in the period from 2008 to 2014, the participants were generally successful in attaining Level 1 from speaking, reading and writing, with an average success rate of 84%, 63% and 84% respectively, this was not the case of listening, in case of which an average success rate equalled to only 49% (see Appendix A for more details).

What helped me to narrow down the scope of the research topic on exploring the link between teaching and testing listening comprehension in the context of L1 courses were STANAG 6001 course feedback questionnaires, in which participants of STANAG 6001 courses are asked to assess the course's effectiveness. The fact that the participants' rating of the effectiveness of L1 courses in developing listening sub-skills targeted by L1 listening test was not as high as one would probably expect suggests that one of the possible reasons for the relatively low average success rate on L1 listening test could

be a lack of link between the listening sub-skills taught and the listening sub-skills tested.

Although the attitudes of second and foreign language researchers and linguists towards different kinds of exam preparatory courses is generally negative, with “teaching to the test” being one of the most frequently stated reasons for holding such attitudes (e.g., Green, 2007; Posner, 2004; Turner, 2004), the attitudes held towards the alignment between what is tested and what is taught are much more positive. In this regard, Popham (2001) distinguishes two kinds of teaching to the test, which he refers to as item-teaching and curriculum teaching. While Popham (2001) is critical of item-teaching, in case of which instruction revolves mainly around practising item types included in the test, he is praiseful of curriculum teaching, which requires a teacher to direct instruction towards a body of knowledge targeted by the test.

Hamilton (2010) holds a similar view, which is evident from the way she argues for the alignment between not only instruction and assessment but among standards, instruction, and assessment. Hamilton (2010) maintains that standards should be the main factor driving “the development of *both* the curriculum and the assessments” (p. 49). Besides Popham (2001) and Hamilton (2010), there are many other advocates of the alignment between testing and teaching (e.g., Alderson & Wall, 1993; Bailey, 1996; Hughes, 2003), who, either implicitly or explicitly, assert that if a test is of high quality and the skills that it measures adequately reflect the skills used in the target language use situation then not only is the alignment between what is tested and what is taught desirable, it is indispensable.

Bearing this in mind, the main purpose of the present study was to shed some light on the degree of alignment between the listening construct taught in L1 courses and the listening construct tested by L1 listening test. To this aim, the study investigated the range of listening sub-skills and strategies developed in L1 courses and tried to compare these to the listening sub-skills outlined in the L1 listening test construct. The study also examined how the participants of L1

courses perceived their preparedness for taking L1 listening test before and after taking it and compared their perception with their performance on the test. This was done in hope that the results of the study will broaden the understanding of teaching and testing listening in the context of the AF SR and help to trigger changes that could eventually lead to the improvement in the alignment between the way listening comprehension at Level 1 is taught and tested.

The dissertation is divided into five chapters. Chapter 1, Introduction, has attempted to provide the rationale for the study. The following chapter, Chapter 2, sets the theoretical context for the study. It begins with reviewing literature of relevance to the research topic, moves on to describe the context in which the study was conducted and finishes with stating the research questions. In the next chapter, Chapter 3, the research methodology used to explore the research questions is presented, with the main focus on describing the research design and characterizing the methods of data collection and data analysis. In Chapter 4, the results of the quantitative and qualitative data analyses are presented and analysed. The limitations of the study as well as main findings and their implications are discussed in Chapter 5. The final chapter, Chapter 6, briefly summarizes the main findings from the study and suggests ideas for future research.

2 Background

Chapter 2 is in three sections. Section 2.1 presents the theoretical context of the study by defining listening, describing contemporary models for listening comprehension, looking in more details into the effect of testing on teaching and reviewing studies of relevance to the research topic. Section 2.2 describes the context in which the study was conducted. Section 2.3 lists the research questions that guided the study.

2.1 Literature review

In the context of learning, teaching and assessing language skills, listening is often considered to be a skill we know least about (e.g., Buck, 1997; Vandergrift & Goh, 2009; Wu, 1998). Our limited understanding of a listening comprehension process is usually attributed to complex thought processes underlying it and also to the fact that listening has long been neglected in second and foreign language (L2/FL) acquisition and research (Gilakjani & Ahmadi, 2011; Harding, Alderson & Brunfaut, 2015; James, 1985; Vandergrift & Goh, 2009). In the late 1990s, Nunan (1997) underlined the ignored importance of listening in L2 learning by calling it “the Cinderella skill” (p. 47). Besides Nunan, there have been many other scholars who have put much effort into showing what key role listening plays in L2/FL learning (e.g., Rubin, 1994; Thompson, 1995; Vandergrift, 2007).

Nowadays, it seems that listening is gradually ceasing to be regarded a passive skill that can be acquired automatically. This trend is perhaps most vividly illustrated by the difference in the way L2/FL listening used to be defined in the past and has been defined more recently. In the past, listening was typically defined in a rather general way, with the emphasis placed on the ability to decode aural input. For instance, Barker (1971) defined listening as "the selective process of attending to, hearing, understanding, and remembering aural symbols" (p.17). Bowen, Madsen and Hilferty (1985) viewed listening as a process of “attending to and interpreting oral language”

(p. 73), with the student's ability "to hear oral speech in English, segment the stream of sounds, group them into lexical and syntactic units (words, phrases, sentences), and understand the message they convey" (p. 73) being a precondition to mastering listening skills. In the same year, Wolvin and Coakley (1985) described listening comprehension as "the process of receiving, attending to, and assigning meaning to aural stimuli" (p. 74).

In more recent years, scholars have become much more specific when defining L2/FL listening, trying to dispel the perception of listening as a simple and passive process. In this regard, Buck (2001) claims that "meaning is not something in the text that the listener has to extract, but is constructed by the listener in an active process of inferencing and hypothesis building" (p. 29). Caldwel (2008) defines listening comprehension in a similar way, calling it "the process of simultaneously extracting and constructing meaning through interaction with oral language" (p. 4). Perhaps one of the broadest definitions of L2 listening is that of Rost (2011), who sees listening as a complex cognitive process involving neurological, linguistic, semantic and pragmatic processing, requiring a listener to take an active role in integrating all the four processing mechanisms. Although varying from each other, the definitions all seem to suggest what Lynch and Mendelsohn (2010) explicitly state: "Listening is an 'active' process, and […] good listeners are just as active when listening as speakers are when speaking" (p. 193).

There are numerous frameworks and descriptive models of listening comprehension attempting to portray the active and complex nature of L2/FL listening comprehension (e.g., Buck, 2001; Field, 2013; Rost, 2011). Most of these frameworks and models are based on perceiving language comprehension as information processing and explore the L2/FL listening process from the perspective of bottom-up and top-down processing.

Bottom-up processing, according to Lynch and Mendelsohn (2010), "involves piecing together the parts of what is being heard in a linear fashion, one by one, in sequence" (p. 197). To put it more explicitly, a bottom-up approach to establishing a theoretical framework of listening comprehension views

listening as a process of decoding acoustic input into phonemes, which are then decoded to construct words, which are further decoded and connected to construct phrases, sentences, and texts. This is followed by a semantic analysis of the content. In the last step of the process, literal meaning is interpreted within the context of a given communication situation (Buck, 2001).

In contrast to bottom-up processing, which relies heavily on processing acoustic input, in top-down processing, listeners make use of their background knowledge, experience and expectations in assigning meaning to what they hear (Lynch & Mendelsohn, 2010). In doing so, they draw upon their schemata, which, as defined by Buck (2001), are "structures for representing knowledge in memory […], including general concepts, situations, events, sequences of events, actions, sequences of actions, etc." (p. 20) and are constantly created and updated (Rost, 2011). According to the bottom-up view of listening comprehension, it is believed that every time we hear something, we activate those of our existing schemata we think may be relevant to understanding the given text (Rost, 2011). Schemata can thus be likened to guides, facilitating the process of listening comprehension.

One of the most influential information processing models, often drawn upon when describing listening comprehension, seems to be Anderson's (2009) language production model, which divides language comprehension into three overlapping phases. The first phase, called perception, involves decoding of the spoken message. In the second, parsing stage, mental representation of the meaning of the words is created. The third stage is utilisation, during which, drawing upon their schemata, the listener completes the interpretation of the spoken message.

Another model based on information processing is Field's (2013) model for L2 listening comprehension, in which listening comprehension is characterized as a five-level process, consisting of input decoding, lexical search, parsing, meaning destruction and discourse representation. While the first three levels are described as lower levels of processing, the last two levels are considered higher levels of processing. It is important to mention here that just like with

Anderson's model, also in case of Field's model, the processing levels do not successively follow each other but run simultaneously.

Buck's (2001) framework for describing listening ability, constructed with an intention to aid test developers in defining listening construct, views listening ability as consisting of linguistic knowledge, to which he refers to as "language competence", and the person's intellectual ability to actually apply the linguistic knowledge, which he terms "strategic competence". Based on the major areas of linguistics, language competence is further sub-divided into grammatical, discourse, pragmatic and sociolinguistic knowledge. The sub-categories of strategic competence include cognitive strategies and meta-cognitive strategies. Cognitive strategies are described as mental processes associated with the processing and storing of linguistic and non-linguistic input and are used to make sense of the input. Meta-cognitive strategies are mental processes managing cognitive processes and deal with assessing, monitoring and evaluating one's listening process and abilities. Similarly to Anderson (2009) and Field (2013), Buck (2001) makes it clear in the description of his framework that its components are not sequenced in strict order, but "interact freely with the acoustic input and with each other to create the interpretation of the text" (p. 29).

Although the three above-mentioned frameworks by no means represent a comprehensive list of all the models and frameworks addressing listening comprehension, they are hopefully sufficient to illustrate two facts about the current view of L2/FL comprehension, with the first being that listening comprehension is an interaction between bottom-up and top-down processing and the second being that all the processes involved in listening comprehension operate simultaneously rather than in any set order.

The realisation of the complexity of listening comprehension has inspired L2/FL acquisition researchers to look for the ways of how to promote L2/FL listening comprehension (e.g., Goh, 2000; Chen, 2013; Fahim & Fakhri Alamdari, 2014), with most of them concluding that in order to successfully

comprehend oral information, a L2/FL listener must have a number of learning strategies at their disposal.

One of the most cited definitions of learning strategies was provided by Chamot (1987), who describes learning strategies as “techniques, approaches, or deliberate actions that students take in order to facilitate the learning and recall of both linguistic and content area information” (p. 71). Based on the categorization proposed by O’Malley and Chamot (1990), learning strategies are usually classified into three main groups, namely meta-cognitive, cognitive and socio-affective strategies.

Meta-cognitive strategies refer to strategies consciously used by a L2/FL listener to plan, monitor and evaluate their comprehension. Examples of meta-cognitive strategies include establishing purpose for listening or monitoring and evaluating comprehension using contexts and prior knowledge (Goh, 2000). Cognitive strategies are related directly to making sense of aural input and are explored from the perspective of bottom-up and top-down strategies (Gilakjani & Ahmadi, 2011). Inferring unfamiliar words based on context or taking notes of key content words are cited by Goh (2000) as typical examples of cognitive strategies. Socio-affective strategies are activities enhancing L2/FL listening comprehension by means of interaction with other people and managing one’s negative emotions. Asking speaker for clarification is used as an example of a social strategy and learning to relax before and during listening exemplifies affective strategies (Goh, 2000).

As already mentioned, the L2/FL learner’s ability to use learning strategies effectively has been recognized by many to be a key factor in facilitating L2/FL listening comprehension. That is perhaps the main reason why it has been highly recommended for the teaching of FL/L2 listening to be strategy-based (e.g., Bagheri & Karami, 2014; Fahim & Fakhri Alamdari, 2014; Mendelsohn, 2006; Vandergrift & Goh, 2009). For instance, Mendelsohn (2006) maintains that strategy-based instruction should “constitute the ‘spinal cord’ or organizing principle on which the listening course should be built” (p. 82).

Besides listening instruction, the deepening understanding of a listening process has also had an impact on how assessing listening is approached these days. Construct validity seems to be the most repeatedly mentioned term in this regard, with many scholars (e.g., Alderson, Clapham & Wall, 1995; Hughes, 2003; Vandergrift & Goh, 2009) highlighting the need for ensuring the test measures the abilities it says it measures. The responsibility placed on test developers to define the tested construct clearly and in detail is even greater now that the potential of language tests to affect what is taught in the classroom has been generally acknowledged (e.g., Alderson et al., 1995; Alderson & Wall, 1993; Hamilton, 2010; Hughes, 2003). This effect of testing on teaching and learning is referred to as backwash and is seen both negatively and positively (Hughes, 2003).

Those perceiving backwash in a negative light claim that tests tend to generate negative effects on teaching and learning, making teachers and students focus more on test preparation then on acquiring language. On the other hand, those viewing backwash in a more positive way believe that tests have a potential to produce positive changes in both teaching and learning practices (Alderson & Wall, 1993) as long as they make teachers direct "instruction towards the body of knowledge or skills that a test represents" (Popham, 2001, p. 16).

Several authors (e.g., Bailey, 1996; Hughes, 2003; Messick, 1996) have attempted to provide practical tips on how to make a test produce beneficial backwash. Meaningful alignment between the abilities tested and the abilities taught seems to be one of the most important factors in achieving beneficial backwash, provided, of course, that the construct is well defined and the test representatively samples the construct being measured. The relationship that should ideally exist between teaching and testing is aptly depicted by Hughes (2003) as "that of partnership" (p. 2).

There have been a number of studies looking into the relationship between teaching and testing. For the purpose of this study, I will, however, focus on

those that have concentrated on investigating the influence of test preparation courses on learning and teaching practices.

In 2009, Gan set out to explore the effectiveness of IELTS (International English language testing system) preparation courses in the context of tertiary institutions in Hong Kong. He did so by examining the differences in exit IELTS test performance between the students who enrolled in IELTS preparation course after they started studying at the university and those who did not. Although no significant difference was found in exit IELTS test scores between the two groups of students, after having compared the mean score of the entry A-level English examination scores between the two groups, Gan (2009) found that the students who later decided to participate in IELTS preparation course were more likely to score lower in the university entrance exam than those who did not. Based on the findings, Gan (2009) concluded that "the experience of taking IELTS […] may be valuable particularly for lower entry A-level students" as it has "the potential of narrowing or closing the gap in English language proficiency between the students" (p. 35).

With the intention of contributing to the research exploring backwash on students' learning, Green (2006) attempted to investigate the impact of teachers' perception of IELTS Academic Writing Module on learners. To this end, Green designed two versions of a student questionnaire, namely a course entry questionnaire and a course exit questionnaire, with the first one intended to investigate the students' expectations of the course at the beginning of the course and the latter one meant to find out the students' perception of the course focus at the end of the course. The two versions of the student questionnaire were distributed to the students from China, who were preparing for university study in the UK by either taking non-IELTS courses (75 learners) or IELTS preparation courses (33 learners). Green also developed a teacher questionnaire, by means of which he intended to find out whether there was a difference in the perception of the course focus between the teachers teaching in IELTS courses (26 teachers) and the teachers teaching in non-IELTS courses (13 teachers). The fact that the questionnaire items were shared across the three different versions of the questionnaire

enabled the researcher to compare the expectations and perceptions of the IELTS learners, the non-IELTS learners and the teachers.

The results revealed that there was no significant difference between the means of IELTS Academic Writing band scores at the course entry and at the course exit between the students attending IELTS preparation courses and non-IELTS courses. The results also showed that the students' expectations concerning the course content were high, with both groups of the students preferring the development of academic writing skills to test-taking skills. As regards the students' assessment of what they had learnt on the course, the majority of skills the IELTS preparation students felt were prioritized during the course were closely related to test preparation. However, skills ranked highest by the non-ILETS students concerned general aspects of academic writing rather than test preparation. The teachers' ratings of the course focus correlated significantly with those awarded by the students, with IELTS courses giving greater focus on the development of skills related to test performance and non-IELTS courses focusing more on the development of writing skills deemed necessary for university study.

Another research study seeming relevant to this thesis is that by Rashidi and Javanmardi (2011), who, in the context of Iran, attempted to examine backwash effect of IELTS test on both learning as well as teaching. As regards learning, the study looked at the impact of IELTS preparation classes on students' learning processes and achievement. As to teaching, it looked at the influence of the IELTS test construct on teaching methods. Based on the results of quantitative and qualitative analysis of the students' answers to a questionnaire, developed to investigate the students' expectations of IELTS preparation course before and after taking it, Rashidi and Javanmardi concluded that IELTS preparation courses influenced the students' learning processes and achievement mostly in a positive way. Qualitative analysis of the data obtained from structured interviews with teachers teaching IELTS courses led to a conclusion that IELTS test and its construct "affected the way of teachers' teaching processes" as "a great amount of time in the course was allocated to making students familiar with the form of the exam" (p. 138)

The studies reviewed above all investigated the backwash effect of high-stakes tests in the context of IELTS assessment and with the focus being on exploring the impact of IELTS preparation courses on students' performance as well as on teaching. Although the studies partly dealt also with the influence of the construct of IELTS test on teaching practices, the issue of alignment between what is tested and what is taught was addressed only marginally or not at all. The main aim of this study is therefore to explore the alignment between the construct that a test measures and instruction in the context of L1 listening test and L1 courses.

2.2 Research context

2.2.1 L1 listening test

L1 listening test forms an integral part of STANAG 6001 exam at Level 1, which is a proficiency exam the main purpose of which is to measure the test-takers' English language proficiency at a given time regardless of their prior preparation. As its name suggests, the exam is designed and the criteria for assessment are set in conformity with STANAG 6001 language proficiency scale. The scale consists of the descriptions of six language proficiency levels, ranging from Level 0 (No proficiency) to Level 5 (Highly-articulate native). In addition to the base six proficiency levels, the scale also comprises descriptions of five plus levels. A plus level is defined by STANAG 6001 (2010) as an indicator of proficiency "that substantially exceeds a 0 through 4 base skill level, but does not fully or consistently meet all of the criteria for the next higher base level" (p. 2). The level of language proficiency achieved is recorded by letters SLP, meaning standardized language profile, followed by a sequence of four digits, representing four language basic skills, sequenced in the following order: listening, speaking, reading and writing (STANAG 6001, 2010).

Although the system based on STANAG 6001 may appear as standardised, adopting the scale and putting into effect the assessment in accordance with

the criteria presented in STANAG 6001 is the responsibility of each NATO member nation (STANAG 6001, 2010). Despite the fact that the Bureau for International Language Cooperation, established in 1996 with the main purpose to promote standardization in the field of language training and testing, is of great support to NATO as well as Partnership for Peace nations in this matter, it is up to each nation alone to develop and implement its own training and testing system (Dubeau, 2006).

In the context of the AF SR, STANAG 6001 exam development, administration and evaluation is in charge of the English Methodology and Testing Branch of the LI. The exam may be taken as a single level exam (from SLP 0 to SLP 1) or as a bi-level exam (from SLP 1+ to SLP 3). The purpose of the single level exam, of which L1 listening test is an inseparable part, is to measure the test-takers' language proficiency in all four language skills in the range from SLP 0 to SLP 1. Figure 1 shows descriptors for listening comprehension at Level 1.

Level 1 – Survival
2. Can understand common familiar phrases and short simple sentences about everyday needs related to personal and survival areas such as minimum courtesy, travel, and workplace requirements when the communication situation is clear and supported by context. Can understand concrete utterances, simple questions and answers, and very simple conversations. Topics include basic needs such as meals, lodging, transportation, time, simple directions and instructions. Even native speakers used to speaking with non-natives must speak slowly and repeat or reword frequently. There are many misunderstandings of both the main idea and supporting facts. Can only understand spoken language from the media or among native speakers if content is completely unambiguous and predictable.

Figure 1. Descriptors for listening comprehension at Level 1 (STANAG 6001, 2010, p. 3)

The construct of L1 listening test is based on the description of typical performance at Level 1 - Survival as designated by STANAG 6001 and also on Buck's (2001) framework for describing listening ability, given in full in Appendix B. It takes up to 25 minutes to complete and contains 20 tasks, which are of three types, namely a multiple choice question, a short answer question and a constructed-response task. The tasks at Level 1 are aimed at testing the test-takers' ability to identify gist, understand a simple main idea as well as to identify specific information and detail. There are maximum two

tasks to one listening passage. The listening passages at Level 1 comprise of a string of short simple sentences and cover basic topics, such as food, accommodation, family, interests, holiday and so on. The types of listening passages include simple telephone messages, announcements, simple conversations, interviews as well as simple narrations, descriptions, instructions and directions. The length of the listening passages is maximum 1.5 minute and they are usually played once only. The following is one example of a listening passage at Level 1 with an accompanying item:

Example 1

Passage	Man:	Excuse me! Is there a newsagent's near here?
	Woman:	Yes, there's one in Church Street next to the bank and there's one in Park Lane opposite the swimming pool.
	Man:	Is that one far?
	Woman:	No, just 2 minutes. That's all.

Item The man is asking for

A a newspaper.
B directions.
C a parking place.
D the time.

As obvious from the examples, all stems and options are given in the target language. As for marking, each correct answer receives one mark. The minimum passing score for Level 1 is 70%. In case of Level 0+, it is 60%. The listening test specifications are provided in full in Appendix C.

2.2.2 L1 language course

L1 course forms an essential part of a language training programme provided by the LI. The programme itself consists of intensive language courses in

compliance with STANAG 6001 at Level 1, Level 2 and Level 3, specialized language courses, basic courses of everyday and military English as well as extensive language courses (http://www.ji.mil.sk).

L1 course focuses on preparing the participants for taking STANAG 6001 Level 1 exam and each participant is obliged to take the exam after completing it. It is held twice a year, once from September till December and once from February till May, at four LI's language training centres in Prešov, Liptovský Mikuláš, Zvolen and Trenčín. The course is intensive, consisting of thirty-eight 45-minute lessons taught weekly for the period of four months. The basic teaching materials used in L1 course include New English File Elementary Book, Essential Grammar in Use, Campaign 1 and Handbook of English Language Proficiency Test Based on STANAG 6001 for Level 1, designed by the English Methodology and Testing Branch of the LI to assist the teachers in preparing L1 course participants for sitting STANAG 6001 exam at Level 1 (http://www.ji.mil.sk).

2.3 Research questions

Three research questions were identified to explore the degree of alignment between the listening sub-skills and strategies taught in L1 courses and the sub-skills and strategies tapped by the tasks included in L1 listening test:

1 To what extent do L1 courses teach listening sub-skills and strategies measured by L1 listening test?
2 How do the participants of L1 courses perceive their preparedness for taking L1 listening test before and after taking it?
3 What is the relationship between the L1 course participants' perception of the effectiveness of L1 courses in developing their listening skills and their performance?

3 Methodology

Chapter 3 describes in detail the research methodology used to explore the research topic. The first sub-section provides information on the research design. The second sub-section identifies the data collection methods utilized to gather data for this study. The chapter concludes with the third sub-section, which presents the methods of data analysis.

3.1 Type of data

In order to arrive at the answers to the research questions, a mixed methods approach was used. There were three main reasons for doing so. First, the combination of qualitative and quantitative data collection methods and data analysis techniques is recommended in social science research (e.g., Cohen, Manion & Morrison, 2007; Dörnyei, 2007; Turner, 2014). Second, as stated by many experts in the field (e.g., Creswell, 2014; Dörnyei, 2007; Turner, 2014), a mixed methods research design allows a better understanding of the phenomenon under study. Third, adopting a mixed methods research methodology makes it possible for a researcher to triangulate the data, allowing the researched topic to be approached from different perspectives and at the same time, it also increases the research outcomes validity (Dörnyei, 2007).

One of the main intentions of this study was to build a picture of how the tested listening construct was taught from the data provided by two main stakeholder groups: students and teachers. For the purpose of not prioritizing any sets of data, a convergent parallel design (Creswell, 2014) consisting of four data collection methods was employed (see Figure 2). In order to obtain the students' viewpoints, a questionnaire was developed. Its pre-test version was distributed to the students prior to and its post-test version was distributed after writing L1 listening test, which was another data collection method used. The teachers' views and opinions were collected through a semi-structured teacher interview. The last research instrument used was a classroom

observation. Its inclusion was motivated by the fact that the study by its very nature explored the relationship between a testing system and classroom practices and employment of a classroom observation is highly recommended in such case (e.g., Alderson & Wall, 1993; Bailey, 1996; Wall, n.d.).

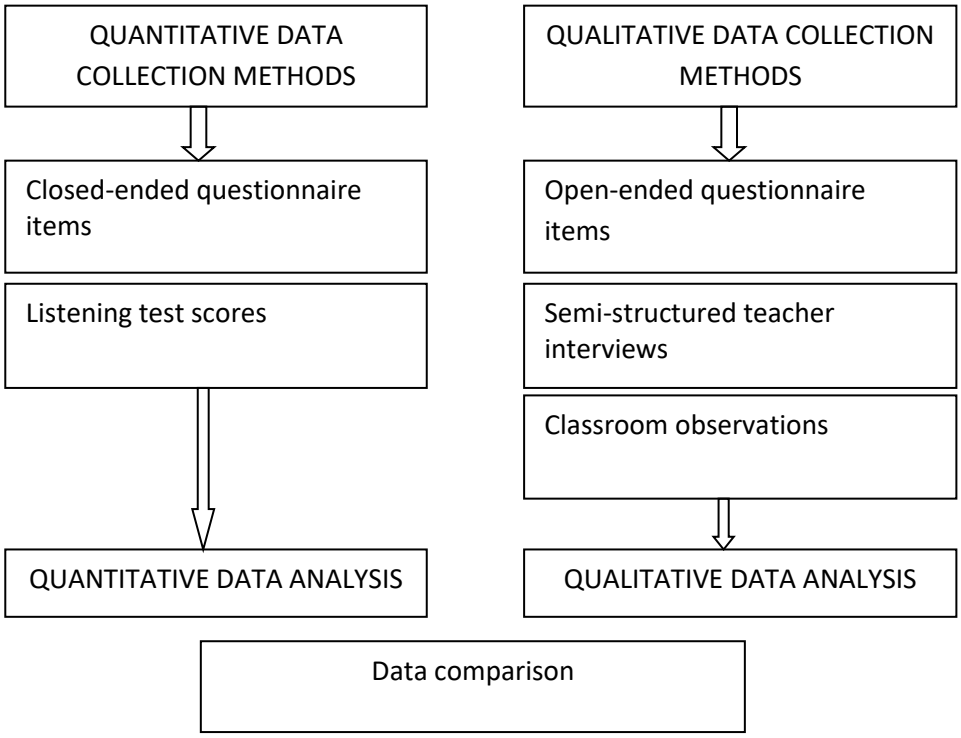


Figure 2. Research design

3.2 Data collection methods

As already mentioned, there were four different research instruments used to collect data, namely (a) questionnaire, (b) L1 listening test, (c) semi-structured teacher interview, (d) classroom observation. The data were collected at the four centres of the LI, where L1 courses were held in the first part of the 2016 academic year. Altogether, four L1 courses, 51 student participants and 8 teacher participants were included in the study.

3.2.1 Questionnaire

The main purpose of the questionnaire was to obtain information about how the student participants perceived the effectiveness of L1 course in developing their language and strategic competence as far as listening was concerned. To this end, two versions of a questionnaire were designed: a pre-test version (Version A) and a post-test version (Version B). Version A, an English translation of which can be found in Appendix D, was intended to give a rough idea of the extent to which listening strategies and sub-skills were developed in L1 courses from the perspective of the student participants. Version B, an English translation of which can be found in Appendix E, was aimed at providing information about how the student participants perceived their ability to actually use the strategies.

The questionnaire design was based on three theoretical frameworks. Since the construct the listening test targets has been defined in the light of Buck's (2001) framework for describing listening ability, it was Buck's framework that guided the questionnaire design. The items generated using this framework were aimed at eliciting information about how the student participants perceived the effectiveness of L1 course in developing their language competence as well as their familiarity with cognitive and meta-cognitive strategies. The framework of language task characteristics, in which, building on Bachman (1990), Palmer and Bachman (1996) describe five characteristics they deem useful when designing test tasks, was also drawn upon. The main aim of the items based on this framework was to learn more profoundly about the student participants' perception of the effectiveness of L1 course in developing their familiarity with the test format. The last framework used was Wenden's (1991) framework of test-wiseness strategies, in which Wenden divides test-wiseness strategies into those used prior to, during and after answering the test. The items developed based on this framework were included into the questionnaire with the main intention to explore how the student participants perceived the effectiveness of L1 course in developing their test-wiseness strategies.

The three theoretical frameworks led to the identification of five content areas the questionnaire was intended to address. These included language competence, cognitive strategies, meta-cognitive strategies, test format familiarity and test-wiseness strategies. When developing questionnaire items, special care was taken to ensure that the items reflecting individual content areas were closely related to the particular theoretical construct underlying them. Research papers by Barta (2010) and Vandergrift (1997) were found very useful in this respect because they include taxonomies of listening comprehension strategies accompanied by examples illustrating their meaning.

After the two versions of the questionnaire were drafted, they were piloted on a group of 12 Level 2 course students. After completing the two questionnaire versions, the pilot participants were asked to explain how they interpreted the meaning of individual items in order to see whether the items were interpreted as intended. Based on the comments made by the pilot participants, the two versions were revised. The revision included a number of changes aimed at simplifying the language used in the questionnaire items. Care was taken to make the questionnaire items easily comprehensible, yet addressing the content areas intended. The revised and near-final versions of the questionnaire were re-piloted on a group of ten Level 3 course students. The re-pilot did not produce any additional changes to the two questionnaire versions.

In their final version, both Version A and Version B consisted of Section I, called Language and Strategic Competence Section, and Section II, called Additional Questions Section. Version A also comprised Section III, General Information Section. The sequence of the items was identical in both versions. The items' wording in Version B was, however, adapted to the after-test context, so while the items in Version A began with "We have been taught to do this and this", the items in Version B began with "I did this and this" or "I was able to do this and this". Table 1 illustrates the questionnaire design.

Table 1

Questionnaire design

Section no.	Section name	Content area	Factor addressed	Question No.
I	Language and Strategic Competence Section	1. Language competence	Grammar knowledge	1, 2, 10
			Discourse knowledge	11, 22
			Sociolinguistic knowledge	23
		2. Cognitive strategies	Inferencing	6, 7
			Elaboration (Imagery; questioning elaboration)	12, 13
			Summarization	27
			Translation	28
		3. Meta-cognitive strategies	Assessing the situation	8, 9
			Monitoring, Self-evaluating	18, 19, 24, 25
		4. Test format familiarity	Characteristics of the test rubrics	3, 14, 15
			Scoring method	4
			Characteristics of the input	29
			Characteristics of the response	30
		5. Test wiseness	Strategies used before answering the questions	5, 16
			Strategies used while answering the questions	17, 20

		Strategies used after answering the questions	21
		Other	26
II	Additional Questions Section	Motivation	31, 32
		Overall effectiveness	33, 34
		Perceived success on Listening test	35, 36
		Listening test fairness	37
		Participants' suggestions	38, 39, 40
III	General Information	Age	41
		Gender	42
		English language learning history	43
		Deployment abroad	44

Section I contained 30 statements judged on a 6-point Likert scale, ranging from "1 – absolutely disagree" to "6 – absolutely agree", with no middle choice included. The forced-choice method was used because at the time the data were collected, the student participants had been attending L1 course for 16 weeks and they were thus expected to have formed an opinion on the research topic. Each of the 30 statements targeted one of the five already specified content areas. As for the statements order, a recommendation made by Dörnyei (2007) was followed and statements targeting the same content were presented in a random order.

Section II included ten additional questions. The answers to seven of these were given on a Likert scale and were aimed at obtaining information about the student participants' motivation to study English, their perception of the overall effectiveness of L1 course in developing their listening skills and their perceived success on L1 listening test. Clarification questions were attached to the answers to these seven questions, where the student participants were

asked to specify the reason for choosing the particular response option. The remaining three questions were open-ended and invited the student participants to provide their own suggestions on how testing and teaching listening comprehension in the context of L1 courses could be improved.

Section III comprised four items and was intended to provide basic information about the student participants' age, gender, English language learning history and deployment abroad.

The two questionnaire versions were administered in paper-and-pencil format to the students attending L1 courses a few days prior to taking STANAG 6001 exam. Altogether, 51 student participants, whose mother language was Slovak, completed the two versions of the questionnaire. As the level of the student participants' English was elementary, the questionnaire was in Slovak. The participants were explained about the research project. They were also informed that the participation was voluntary and that all the data collected would be anonymised. All the participants signed a consent form before the data collection began.

The sample consisted of 51 participants, out of which 42 (82.4%) were males and 9 (17.6%) were females. The majority of the participants surveyed (78.5%) were between 31 and 40 years old. 74.5% of the participants reported they studied English before they enrolled in L1 course. Over half of these (52.6%) studied English for over four years. However, 92% of those studying English noted they finished studying English more than four years ago. Almost 30% of all the participants (27.5%) reported they were deployed on a mission abroad (see Appendix F for more details).

3.2.2 L1 listening test

In order to answer the third research question of whether there was a correlation between the student participants' perceived effectiveness of L1 course in developing their listening skills and their performance on L1 listening

test, it was necessary to measure the student participants' listening comprehension and to obtain their total score. An actual L1 listening test was used to that end.

The test, administered to the same 51 student participants who filled in the two questionnaire versions, was taken by the student participants immediately after they completed Version A of the questionnaire. It took 29 minutes to complete and consisted of 20 items aimed at assessing the student participants' listening abilities, in particular their ability to identify specific information, important details, simple main idea and gist in texts at Level 1. Nine of the items were multiple-choice items, six were short answer questions and the remaining five were constructed-response items. There were 16 listening passages in the test, with all of them but two played once only.

3.2.3 Semi-structured teacher interview

The main objective of a semi-structured teacher interview was to provide the teachers teaching L1 courses a chance to express their attitudes towards teaching and testing listening comprehension at Level 1. All the interviews were conducted in the teacher participants' native language so as to ensure the teachers felt comfortable enough to share their views. It was believed that the data collected using this research method would shed some light on how students' listening sub-skills and strategies are developed in L1 courses. The interviews were based on a semi-structured interview guide (for its English translation, see Appendix G) and were constructed following the suggestions made by Dörnyei (2007) and Cohen et al. (2007).

The interview guide comprised four sections. Section I provided information about the interview purpose and structure and it also reassured participants about the confidentiality of their responses and their anonymity. Section II included questions aimed at eliciting information about the teacher participants' experience with teaching English, working for the LI and teaching L1 courses. Section III comprised a collection of questions covering five

content areas, including teacher's usual way of teaching listening comprehension, their training in teaching listening comprehension, challenges involved in teaching listening comprehension, their opinion on testing listening according to STANAG 6001 and their perceived reasons for low success rate in L1 listening test. Section 4 encouraged the teacher participants to further comment on the issue of teaching listening skills and strategies in L1 courses.

The interview guide was piloted in one, about 20-minute long, interview with a teacher working for the LI, who had been working for the LI for 3 years and had experience with teaching L1 courses but was not involved in the study. The pilot led to several minor changes in the questions' wording. Before the interviews were carried out, the researcher contacted the teacher participants to inform them about the purpose of the research and to agree on a date for an appointment. The participants were also asked to sign an informed consent, in which they were assured about the confidentiality of the data provided.

The data collection took place at the four centres of the LI over three weeks, from May 16, 2016 to June 3, 2016. The teacher participants included eight teachers involved in teaching L1 courses in the first half of the 2016 academic year, two of which were males and six were females. The participants were between 35 and 62 years old, with a mean age of 48.50 (SD=11.81). Their experience with teaching English was rather rich, ranging from 8 to 36 years (mean=21.50, SD=11.43). The participants' work experience at the LI ranged from 1 year to 17 years, with a mean of 9.75 (SD=6.52). The number of L1 courses they had taught ranged from 1 to 20 (mean=7.88, SD=7.02).

3.2.4 Classroom observation

As research focusing on exploring the link between a test and classroom practices has been criticised for relying too much on student or teacher participants' accounts of what happens in the classroom instead of going to the classroom and seeing what really happens there (Alderson & Wall, 1993),

a semi-structured classroom observation was employed to triangulate the data obtained from the questionnaires and the interviews. Its main aim was to find out how listening skills and strategies tapped by L1 listening test were developed during listening-focused classes. A semi-structured observation schedule, given in Appendix H, was developed for this purpose. Its design comprised five development phases, including design, pilot, revision, re-pilot and classroom observation administration.

A simple observation schedule, which enabled the researcher to take notes of classroom events in real time, was created. It contained six column sections focusing on different aspects of a listening-focused class, including type of activity, its duration, students' grouping format and students' level of interest. The section named Notes was used to record detailed notes of how teaching listening was approached. There also was a Post-observation data analysis section included, designed to aid in a post-hoc analysis of the data collected.

Once drafted, the observation schedule was piloted with 12 Level 2 course students for a period of two lessons, totalling 90 minutes. The pilot resulted in several changes to the schedule's format, for example using A3 size paper rather than A4 one, due to a large amount of data to be recorded.

Before the observations were carried out, the teachers whose classes were to be observed were contacted to be informed about the purpose of the research and to schedule each of the classroom observations. The teachers were at the same time assured that the observations were not intended to evaluate their performance but to potentially improve alignment between the listening sub-skills tested and taught. Both teachers and students were asked to confirm their participation in the study by signing a consent form.

There were altogether ten 45-minute long classes, totalling to 7.5 hours, observed at the four centres over a period of three weeks in May 2016. This included four Level 1 groups, 51 students and 8 teachers. As regards the student participants, they were the same as those who participated in the questionnaire. The teacher participants were the same as those who took part

in the teacher interviews. Each group was observed on two or three occasions.

3.3 Methods of data analysis

3.3.1 Questionnaire

Since the data obtained from the two questionnaire versions were hoped to aid in answering all the three research questions, their analysis, carried out by a statistical package SPSS 21.0, involved several steps.

As regards Section I (Item No.1 – No. 30), the reliability of the questionnaire scales was first assessed by calculating Cronbach's alpha values for each questionnaire version. Following this, Section I items were grouped into five sub-scales according to the five content areas they were meant to address and the reliability of each of these five sub-scales was computed for both Version A and Version B. By looking at Cronbach's alpha if item deleted value, inter-item as well as item-total correlations, the researcher then tried to identify the items which were not contributing to the reliability of individual sub-scales.

After that, missing data was checked for, but there were no cases with missing data found. To indentify multivariate outliers, Mahalanobis distance was calculated, using the linear regression procedure.

Since data obtained from Likert items are considered to be ordinal (Allen & Seaman, 2007; Subedi, 2016), medians and modes for each of the items included in both questionnaire versions were calculated to see how the student participants perceived the L1 courses' effectiveness in preparing them for taking L1 listening test. The mean of means and standard deviations for the five sub-scales of Version A and Version B were calculated to make the comparison of the pre-test and post-test means clearer. The main reason for treating the sub-scales as interval instead of ordinal was Brown's (2011) argumentation that "Likert scales contain multiple items and can be taken to

be interval scales“ (p. 13), with „intervalness“ being „an attribute of the data, not of the labels“ (p. 11).

For the purpose of determining whether parametric or non-parametric analyses were needed to investigate if there were some statistically significant differences in the pre-test and post-test means, the data were first checked for normality. This was done by using Kolmogorov-Smirnov test and also by dividing skewness and kurtosis statistics for each sub-scale by their standard error. As all the Version A and Version B sub-scales were found to be reasonably normally distributed, a paired samples test was used to determine the significance of differences in the pre-test and post-test means. Since in case of multiple testing the probability of making type I error increases (Curtin & Schultz, 1998), Holm-Bonferroni correction procedure was applied to avoid rejecting a null hypothesis when it is true (Bachman, 2004). The alpha level was set to $p<.01$ (.05/5).

As for Section II, each variable included in the section was first checked for missing values. Several missing values were identified. To avoid shortcomings related to listwise deletion, such as sample size reduction (Allison, 2001), pairwise deletion was used. Since pairwise deletion assumes that missing data are missing completely at random (Allison, 2001), Little’s Missing Completely at Random Test analysis was utilised to test a null hypothesis that the data were missing completely at random.

Following this, the variables were checked for outliers by means of visual inspection of histograms and box plots.

Measures of central tendency and dispersion for each questionnaire item included in Section II (items No. 31- No. 37) were then calculated. This applied to both versions of the questionnaire. Before statistical significance of the difference between the pre-test and post-test variables was assessed, the distributions of all the variables were evaluated for normality by dividing the values of skewness and kurtosis by their respective errors and also by using Kolmogorov-Smirnov test.

As the variables failed to meet the normality assumption, the Wilcoxon signed-rank test was used to explore whether the experience of sitting the test changed something about the student participants' motivation and the way they perceived the overall effectiveness of L1 courses in developing their listening skills, their prospect of success on the test and the test fairness.

In an attempt to investigate the relationship between the student participants' overall satisfaction with the preparation for taking L1 listening test (question 33) and their performance, Spearman rank correlation was used. The main reason for choosing this correlation technique was that preliminary analyses showed one of the two variables not to be normally distributed.

In analysing the responses to seven clarification questions and three open-ended questions included in Section II, a four-stage process of a qualitative content analysis proposed by Bengtsson (2016) was adhered to. In the first stage, the responses were read and re-read and their content was searched for meaning units of relevance to the research topic. These were then labelled a code. The second stage involved re-reading the responses for the purpose of ensuring no relevant information was left out. The third stage included sorting the codes into categories and sub-categories. In the last stage, the categories and its individual sub-categories were interpreted both qualitatively and quantitatively. In order to ensure the content analysis was trustworthy, the first three stages of the content analysis process were repeated three times.

In case of Section III, descriptive statistics was run on all the variables and was used for describing the student participants sample group.

3.3.2 L1 listening test

The students' responses on the test items were scored by two raters, the researcher and a member of the language testing team. In order to see how the test performed, the normality of the test scores' distribution was first evaluated by means of analyzing the findings from the descriptive statistics.

Cronbach's alpha was then employed to examine the test's internal consistency. The performance of individual items was examined by looking at their facility value, discrimination index and the value of Cronbach's alpha if item deleted.

3.3.3 Semi-structured teacher interview

The interview data were analysed through a thematic analysis as this data analysis method is believed to enable a researcher to capture the meaning research participants attribute to a researched phenomenon (Riger & Sigurvinsdottir, 2016). Riger and Sigurvinsdottir (2016) define thematic analysis as "a method for analysing qualitative data that involves searching for recurring ideas (referred to as *themes*) in a data set" (p. 33). Braun and Clarke (2006) provide a detailed guide consisting of six phases a researcher should go through when performing a thematic analysis.

Following the guide, the researcher first immersed herself in the data. This involved transcribing the interviews, re-reading the transcripts and making notes of potential codes. In the second phase, the transcripts were manually coded. As the data were coded with the focus being on how the listening construct tested by L1 listening test is taught, coding was theory-driven. After this, the codes were grouped into potential themes, which were then visualized by means of an initial thematic map. To ascertain that the codes grouped under the same theme fitted together and the potential themes reflected the data set adequately, the data set was re-read and the themes were reviewed and refined in the next phase. This phase finished with the creation of a final thematic map, depicting the themes and sub-themes and their interrelationships. In the penultimate phase, the central idea of each theme was identified and named so that it described the theme fittingly. As suggested by Braun and Clarke (2006), a detailed analysis of each theme, which was used as a foundation for further themes analysis and interpretation, was written at this point. The last phase included final analysis, presenting the findings and their interpretation in the Results section.

Throughout the process of performing the thematic analysis and also in reporting the findings, the researcher strived for neutrality. The interviews were recorded in Slovak (see Appendix I for the English translation of one of the teacher interview transcripts), and the extracts used to illustrate the themes and key points of discussion were translated into English. As the analysis was conducted by one researcher only, to ensure a certain level of quality, the whole process was repeated twice.

3.3.4 Classroom observation

To examine whether there was alignment between the listening construct taught in L1 courses and the one tested by L1 listening test, the observation schedules were analysed in light of Buck's (2001) framework for describing listening ability. Since student performance is believed to be influenced by student's familiarity with the test content and format as well as by student's awareness of test-taking strategies, Palmer's and Bachman (1996) framework of language task characteristics and Wenden's (1991) framework of test-wiseness strategies were also drawn upon.

The observation data analysis was carried out following the same six-stage qualitative thematic analysis that was used in the analysis of the data obtained from the teacher interviews and was described in detail in Section 3.3.3. Appendix J includes a scan of one of the completed observation schedules.

4 Results and their analysis

In Chapter 4, the results of the quantitative and qualitative data analyses are presented and interpreted. As the results are presented for each of the four data collection instruments separately, the chapter is sub-divided into four sub-sections. The first sub-section presents the results of the analyses of the data obtained from the questionnaires. The second sub-section describes the results of the analyses of the data gathered from L1 listening test. The third sub-section provides the results based on the qualitative analysis of the data collected through the teacher interviews. The last sub-section presents the results based on the analysis of the data obtained through the classroom observations.

4.1 Questionnaire

4.1.1 Section I

Table 2 shows the values of Cronbach's alpha for Section I scales and also for each of the five sub-scales of Version A and Version B. The Cronbach's alpha values equalled to 0.87 and 0.95 respectively, which, according to Pallant (2007) indicates a very good internal consistency.

The values of Cronbach's alphas for the five sub-scales of both questionnaire versions ranged from 0.66 to 0.90, which indicates acceptable to high internal consistency (Pallant, 2007). Relatively high inter-correlations among the items representing individual content areas (see Appendix K) suggest that the items could be said to have reflected the same construct, which, as believed by Trochim (2002), could be considered the evidence of construct validity of the individual sub-scales. The only item which did not contribute positively to the internal consistency of its respective sub-scale and whose item-total correlation was negative was Item 8 (Version B), which was therefore deleted from both versions of the questionnaire. After its deletion, Cronbach's alpha of the meta-cognitive sub-scale of Version B increased from 0.66 to 0.81.

Table 2

Cronbach's alpha coefficient for Version A and Version B scales and their sub-scales

	Version A	**Version B**
	α	**α**
Section I	0.87	0.95
Content area I - Language competence	0.77	0.90
Content area II – Cognitive strategies	0.70	0.80
Content area III – Meta-cognitive strategies	0.82	0.66
Content area IV – Task characteristics	0.80	0.87
Content area V – Test-wiseness	0.74	0.77

Based on the results of Mahalanobis distance analysis, which was used to detect multivariate outliers, four outliers with $p<.001$ were excluded from further analysis.

Table 3 summarizes the measures of central tendency and dispersion for items on Version A and Version B, arranged according to the five sub-scales. It can be seen from the table that most of the medians hovered around 4.00 and 5.00, which means that the student participants generally agreed or partly agreed with the questionnaire statements about the effectiveness of L1 course in developing their listening sub-skills and strategies both before and also after taking the test.

In order to make the comparison between the pre-test and post-test medians and means clearer, the measures of central tendency and dispersion for each sub-scale were calculated for both questionnaire versions (see Table 4).

Table 3

Pre-test and post-test measures of central tendency and dispersion for Section I questions

	Pre-test						**Post-test**					
Language competence (Sub-scale 1)												
	Q1	**Q2**	**Q10**	**Q11**	**Q22**	**Q23**	**Q1**	**Q2**	**Q10**	**Q11**	**Q22**	**Q23**
Mean	4.10	3.85	3.68	3.77	4.28	3.81	4.00	3.85	4.09	3.96	4.45	3.79
Median	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	5.00	4.00
Mode	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	5.00	4.00
SD	.70	.78	.78	.76	.83	.97	1.00	.88	.91	.78	1.12	.93
Range	3.00	3.00	3.00	4.00	5.00	4.00	5.00	5.00	4.00	3.00	5.00	5.00
Cognitive strategies (Sub-scale 2)												
	Q6	**Q7**	**Q12**	**Q13**	**Q27**	**Q28**	**Q6**	**Q7**	**Q12**	**Q13**	**Q27**	**Q28**
Mean	4.79	3.83	4.17	4.04	3.68	3.40	4.74	3.55	4.38	4.17	3.89	3.92
Median	5.00	4.00	4.00	4.00	4.00	4.00	5.00	4.00	4.00	4.00	4.00	4.00
Mode	5.00	4.00	4.00	4.00	4.00	4.00	5.00	4.00	5.00	5.00	4.00	4.00
SD	.98	.82	1.01	.86	1.13	1.17	.87	1.00	.92	.89	1.07	1.25
Range	3.00	4.00	4.00	4.00	5.00	5.00	3.00	5.00	4.00	3.00	5.00	5.00
Meta-cognitive strategies (Sub—scale 3)												
	Q9	**Q18**	**Q19**	**Q24**	**Q25**		**Q9**	**Q18**	**Q19**	**Q24**	**Q25**	
Mean	4.70	4.81	4.51	5.17	4.64		4.85	4.49	4.30	4.04	4.09	
Median	5.00	5.00	5.00	5.00	5.00		5.00	4.00	4.00	4.00	4.00	
Mode	5.00	5.00	5.00	5.00	5.00		5.00	4.00	4.00	4.00	4.00	
SD	1.12	.86	.93	.79	1.03		.86	.86	1.08	1.02	.93	
Range	5.00	3.00	3.00	3.00	3.00		4.00	3.00	4.00	5.00	5.00	

Task format familiarity (Sub-scale 4)

	Q3	Q4	Q14	Q15	Q29	Q30	Q3	Q4	Q14	Q15	Q29	Q30
Mean	3.15	3.89	4.87	4.79	3.02	3.96	4.09	4.26	4.53	4.53	4.21	4.26
Median	3.00	4.00	5.00	5.00	3.00	4.00	5.00	5.00	5.00	5.00	4.00	4.00
Mode	2.00	4.00[a]	5.00	5.00	3.00	4.00	5.00	5.00	4.00[a]	5.00	4.00	4.00
SD	1.85	1.76	1.06	1.04	1.13	1.23	1.49	1.24	1.08	.98	.91	.94
Range	5.00	5.00	4.00	4.00	4.00	5.00	5.00	5.00	4.00	4.00	4.00	4.00

Test-wiseness (Sub-scale 5)

	Q5	Q16	Q17	Q20	Q21	Q26	Q5	Q16	Q17	Q20	Q21	Q26
Mean	5.79	5.28	5.28	4.43	3.94	4.57	5.09	4.75	4.87	4.45	4.11	3.77
Median	6.00	5.00	5.00	5.00	4.00	5.00	5.00	5.00	5.00	4.00	4.00	4.00
Mode	6.00	5.00	6.00	5.00	3.00	5.00	5.00[a]	5.00	6.00	4.00	4.00[a]	4.00
SD	.46	.74	.88	1.12	1.10	.85	.91	.97	1.10	.97	1.31	.89
Range	2.00	3.00	4.00	4.00	4.00	3.00	3.00	3.00	4.00	4.00	4.00	5.00

Table 4 shows that the pre-test and post-test means ranged from 3.92 to 4.88. For the purposes of the questionnaire results analysis, it was decided to refer to the mean of between 3.00 and 3.99 as representing students' low satisfaction and the mean of between 4.00 and 4.99 as representing students' moderate satisfaction with the way individual areas of their listening ability was developed in L1 course.

From the five facets addressed by the two versions of the questionnaire, the student participants expressed the weakest satisfaction with the development of their language competence (pre-test mean=3.92; post-test mean=4.02). Prior to taking the test, the student participants were also not quite satisfied with their test format familiarity (pre-test mean=3.95). However, an increase in the post-test mean (4.31) indicates that the student participants' familiarity with the test format was better than they thought it would be. The last facet in case

of which the students expressed weak satisfaction was cognitive strategies (pre-test mean=3.99). Although, similarly to test format familiarity, there was an increase in the post-test mean (4.11), the increase was only slight.

Table 4

Measures of central tendency and dispersion of Version A and Version 2 sub-scales

Pre-test					**Post-test**				
Mean	**Median**	**Mode**	**SD**	**Range**	**Mean**	**Median**	**Mode**	**SD**	**Range**
Language competence									
3.92	4.00	4.00	.55	2.83	**4.02**	4.00	3.50	.74	3.33
Cognitive strategies									
3.99	4.00	3.83[a]	.62	2.50	**4.11**	4.17	4.17	.63	2.83
Meta-cognitive strategies									
4.77	4.60	4.60	.72	2.60	**4.35**	4.40	4.20	.65	2.80
Test format familiarity									
3.95	3.83	3.83	.91	3.83	**4.31**	4.50	4.00[a]	.81	3.17
Test-wiseness									
4.88	5.00	5.00	.56	2.00	**4.50**	4.50	4.67	.64	2.83

In contrast to the student participants' relatively weak satisfaction with the effectiveness of L1 course in the development of their language competence, test format familiarity and cognitive strategies, the student participants expressed moderate satisfaction with the development of their meta-cognitive and test-wiseness strategies, with the pre-test means equalling to 4.77 and 4.88 respectively. However, the experience of sitting the test led to a decrease in the means of both meta-cognitive (post-test mean=4.35) and test-wiseness strategies (post-test mean=4.50).

Before finding out whether the differences between the pre-test and post-test means were statistically significant, the distribution of the scores for individual sub-scales was first checked for normality. The values resulting from the division of skewness and kurtosis by their respective standard errors all fell within the reasonably normal distribution range of ±2 (Brown, 2004; Green, 2013), suggesting all the sub-scales were normally distributed (see Appendix L). However, the results of Kolmogorov-Smirnov test showed that in cases of two sub-scales, namely Version A Cognitive strategies sub-scale and Version B Test-wiseness sub-scale, the Sig. value was less than 0.05 (see Appendix M), indicating that the null hypothesis of a distribution being a normal distribution should be rejected (Pallant, 2007). However, in case a sample has more than 30 cases, contains about the same number of cases and does not have any outliers, a parametric test can be run on a data set even if it fails to meet the assumption of normality (Green, 2013). Since the two data sets met all of the three above assumptions, it was decided to run a parametric test on them. The results of the paired-samples t-tests showed that at p-value less than .01, there was a statistically significant difference between the pre-test and post-test means of the meta-cognitive scale and test-wiseness scale (see Appendix N).

As regards meta-cognitive scale, there was a significant decrease in the means before (M=23.83, SD=3.60) and after (M=21.77, SD=3.24) writing the test, $t(46)=2.93$, $p<.01$ (two-tailed). The mean decrease was 2.06, with a 95% confidence interval ranging from .65 to 3.48. The eta squared statistic (.16) indicated a large effect (Cohen, 1988).

As for the test-wiseness scale, there was a significant decrease in the means before (M=29.28, SD=3.34) and after (M=26.36, SD=3.76) sitting the test, $t(46)=4.35$, $p<.01$ (two-tailed). The mean decrease was 2.92, with a 95% confidence interval ranging from 1.57 to 4.26. The eta squared value (.29) indicated a large effect (Cohen, 1988), with a substantial difference in the test-wiseness ratings obtained before and after sitting the test.

4.1.2 Section II

Table 5 shows the number and percentage of missing values per each variable included in Section II of both questionnaire versions.

Table 5

Number of missing values per each question included in Section II

		Version A		Version B	
Variable	**Question No.**	**#NAs*/%**		**#NAs*/%**	
Motivation	31.	0	0	1	2
	32.	0	0	1	2
Perceived overall effectiveness of L1 courses in developing listening skills	33.	1	2	4	7.8
	34.	1	2	5	9.8
Perceived prospects of success on L1 listening test	35.	0	0	0	0
	36.	3	5.9	3	5.9
Perceived fairness of L1 listening test	37.	1	2	6	11.8

*#NAs = number of missing values

To test the null hypothesis that the missing data are missing completely at random, Little's Missing Completely at Random (MCAR) test was conducted for all the items included in Version A and Version B. Since the Little's MCAR test was found not to be statistically significant in case of both Version A ($x^2(11)=13.60$, p=.256) as well as Version B ($x^2(31)=34.57$, p=.301), the null hypothesis was retained, confirming that pairwise deletion of the missing data could be used.

Visual inspection of the histograms and box plots depicting the distribution of ratings on the questions included in Section II of Version A and Version B revealed several outliers. However, since in case of each question, the 5% trimmed mean value was very similar to the mean value, based on a recommendation made by Pallant (2007), it was decided to include the cases in all subsequent analyses.

Table 6 depicts measures of central tendency and dispersion for each questionnaire item included in Section II for both questionnaire versions.

To explore whether the differences in the pre-test and post-test medians were statistically significant, a Wilcoxon signed-rank test was opted for because the variables failed to meet the normality assumption (see Appendix O and Appendix P). The Wilcoxon test showed that the experience of taking the test did not have an impact on the way the student participants perceived the importance of passing the test ($p=.48$) and the overall effectiveness of the Level 1 course in developing their listening ability ($p=.85$). It also seemed to have no effect on their evaluation of the sufficiency of time devoted to developing listening sub-skills ($p=.08$), on their perceived prospect of success on the test ($p=.32$) as well as on their evaluation of the test fairness ($p=.48$).

The results, however, showed that there was a statistically significant decrease ($p=.02$) in the student participants' perception of their motivation to learn English before (median=6) and after (median=5) writing the test ($Z=-2.333$, $p<.05$, $r=.23$) and there also was a significant increase in the way the student participants perceived the test difficulty ($Z=3.682$, $p<.05$, $r=0.37$), suggesting the student participants perceived the test to be more difficult before than after sitting it. The median increase was 1.0.

The results of the Spearman rank correlation showed that the student participants' overall perception of the effectiveness of L1 course in preparing them for taking L1 listening test and their actual performance was not correlated ($r=.12$, $n=47$, $p>.05$).

Table 6

Pre-test and post-test measures of central tendency and dispersion for Section II questions

Pre-test					**Post-test**				
Mean	**Median**	**Mode**	**SD**	**Range**	**Mean**	**Median**	**Mode**	**SD**	**Range**
Q31 Motivation									
5.35	6.00	6.00	.80	3.00	5.22	5.00	5.00	.79	3.00
Q32 Motivation									
4.90	5.00	5.00	.70	3.00	4.94	5.00	5.00	.74	4.00
Q33 Perceived overall effectiveness of L1 courses in developing listening skills									
4.42	4.00	4.00	.84	3.00	4.45	4.00	4.00	.93	4.00
Q34 Perceived overall effectiveness of L1 courses in developing listening skills									
1.24	1.00	1.00	.43	1.00	1.30	1.00	1.00	.47	1.00
Q35 Perceived prospects of success on L1 listening test									
2.33	2.00	2.00	.93	3.00	2.82	3.00	3.00	.89	3.00
Q36 Perceived prospects of success on L1 listening test									
1.38	1.00	1.00	.49	1.00	1.46	1.00	1.00	.50	1.00
Q37 Perceived fairness of L1 listening test									
1.56	2.00	2.00	.50	1.00	1.49	1.00	1.00	.51	1.00

As illustrated in Table 7, three main categories emerged from the content analysis of the responses to the clarification and open-ended questions, namely (a) appreciation of the way listening is taught, (b) objections on the listening instruction and (c) factors limiting the effectiveness of teaching listening. Each of the categories consisted of five to six subcategories.

Table 7

Summary of comments on the efficiency of L1 courses in the development of listening skills

Meaning unit	Subcategory	No.	%	Category
"The teachers do their best when teaching listening"	Teacher's approach to teaching listening	21	41	1. Appreciation of the way listening is taught
"When practising listening comprehension, listening passages are played twice or more times and everything becomes much clearer."	Repeated listening	5	10	
"Listening with evaluation has been very useful for me"	Encouraging the development of listening strategies	5	10	
"The course has been efficient in developing my listening skills because we have been listening to recordings with different accents"	Providing recordings with a variety of accents	3	6	
"The course has been efficient in improving my listening comprehension because I learnt a lot of new words"	Vocabulary boost	1	2	
"It would be good if the course contained more lessons focused on practising listening comprehension"	The need of devoting more time to practising listening	25	49	2. Objections

"I would suggest to practice listening more and to devote more time to thinking before moving on to another exercise."	Insufficient development of meta-cognitive strategies	8	16
"I think we should have started developing our listening skills from listening to the most simple words and phrases and then gradually move on."	Need of gradual development of listening skills	5	10
"I have no information about the exam."	Lack of information provided about the exam	5	10
"Listening should be taught from different sources and the teachers should adjust to our needs and not only blindly follow the curriculum like in the era of the Czechoslovak Socialist Republic."	Need of more varied recordings	4	8
"In case of beginners, it is necessary to begin from the very basics."	Need of adapting listening instruction to weaker students	2	4
"Sometimes it seems to us that the difficulty level of listening passages is higher than Level 1."	Problem with teachers' ability to choose appropriate-level texts	2	4
"We have practiced writing more than listening."	Other skills preferred	2	4

"It would be good to equip the classrooms with better technical equipment, such as bigger white boards, new CD players and mainly recordings of higher quality"	Obsolete technical equipment	15	29
"The course should last longer."	Short course duration	9	18
"Fewer students in the classroom."	Too many students in the classroom	1	2
"Provide the students with more opportunities to talk to native speakers."	Limited allotment of lessons with native speaker lecturers	1	2

The first category depicts the aspects the student participants liked about listening instruction. The second category describes the elements of listening instruction the student participants perceived negatively. The last category refers to material, technical and financial constraints, which were thought by the student participants to obstruct teachers from delivering an effective listening lesson.

4.2 L1 listening test

The results yielded from the descriptive statistics (Table 8) show that the measures of central tendency were relatively far from each other, with the mode and the median having higher values than the mean, which, according to Green (2013), suggest that the test scores were negatively skewed. There were three other indicators of the test scores distribution having a negative skew. First, two standard deviations could fit below the mean while only one could fit above it (Green, 2013). Second, the value resulting from dividing the value of skewness by its standard error fell outside the ±2 range (Bachman,

2004; Carr, 2011). Third, the scores were clustered at the right side of the histogram (see Figure 3). All this suggested the test was relatively easy for the students. With a cut-off score set at 14 points (70%), 57% of the student participants passed the test, which may seem good enough, but it is seen as unsatisfactory in comparison to, for example, 84% averaged success rate on the Level 1 test of writing.

Table 8

The results of descriptive statistics for Level 1 listening test

Mean	12.98
Median	14.00
Mode	17.00
Std. Deviation	4.60
Skewness	-1.03
Std. Error of Skewness	.33
Kurtosis	.33
Std. Error of Kurtosis	.66
Range	19.00

The positive value of kurtosis means the distribution of the scores was slightly leptokurtic, suggesting the students performed similarly on the test (Green, 2013). However, the value obtained from dividing the value of kurtosis by its standard error (.51) fell within the ±2 range, meaning the distribution was reasonably normally distributed as far as its density was concerned (Bachman, 2004; Carr, 2011).

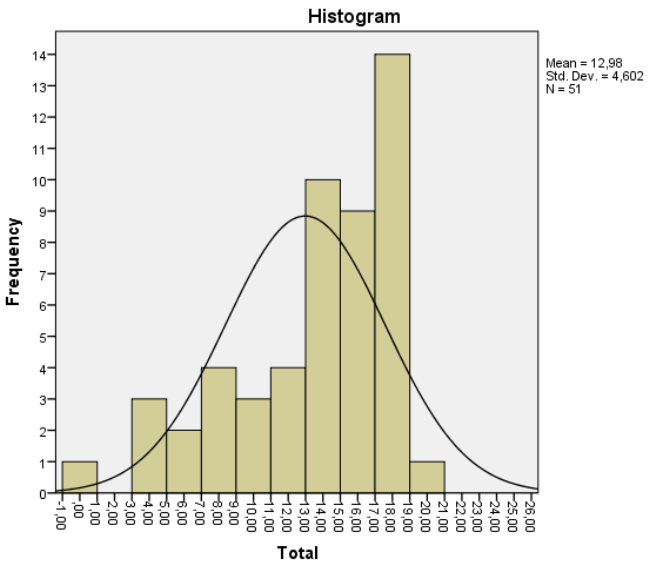
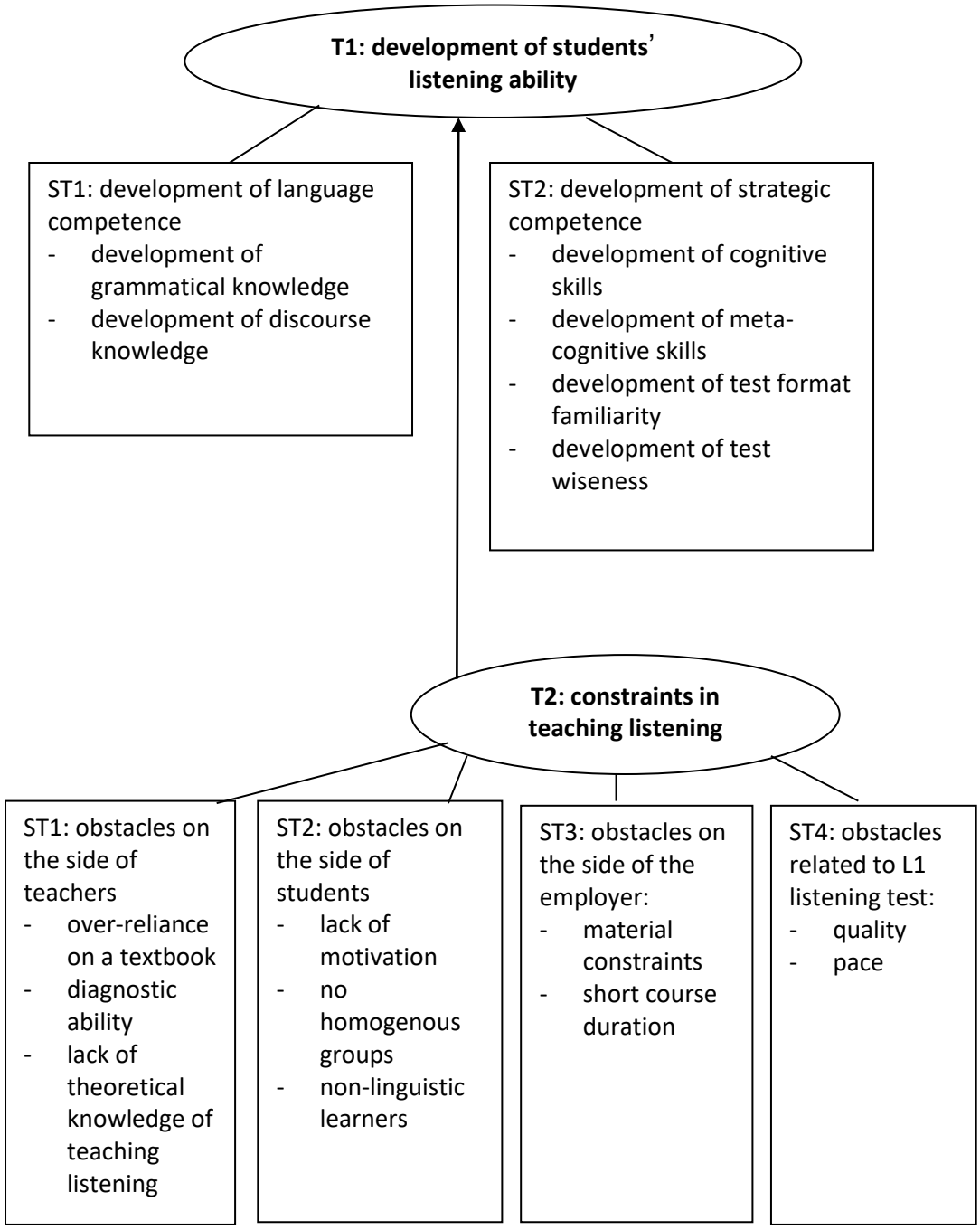


Figure 3. Distribution of L1 listening test scores

The Cronbach's alpha of the test was .86, which is considered to be sufficient (Carr, 2011; Cohen et al., 2007; Hughes, 2003; Rowntree, 1981). The item facility values ranged from 31% to 96% and the discrimination indices varied from .28 to .76. There was no item, the deletion of which would increase the value of Cronbach's value (see Appendix P).

4.3 Semi-structured teacher interview

The length of the teacher interviews ranged from 13 to 25 minutes and contained from 960 to 2307 words. The thematic analysis led to the identification of 31 codes which were grouped into two main themes related to the way the listening construct tested is taught. These included: (a) development of students' listening ability; and (b) constraints in teaching listening. The first theme was further subdivided into two sub-themes and six sub-sub-themes. For the second theme, five sub-themes and ten sub-sub-themes were identified. A final thematic map presented in Figure 4 shows the final themes and sub-themes and the interrelations among them.



Note: T = theme, ST = sub-theme

Figure 4. Themes and sub-themes that emerged from the thematic analysis of the teacher interviews

4.3.1 Theme 1: development of students' listening ability

When talking about teaching listening comprehension, all the teachers, whether openly or implicitly, talked about a significant role the development of students' language as well as strategic competence should play in developing students' listening ability.

4.3.1.1 Subtheme 1: development of language competence

When asked to describe how they taught listening, many of the teachers provided a relatively specific description of how they developed their students' language competence. Most of their remarks were related to the development of grammatical and discourse knowledge.

As regards the development of grammatical knowledge, the teachers mainly talked about the importance of teaching lexis and pronunciation. With respect to teaching lexis, the majority of the teachers explained how they taught vocabulary before and after listening. As far as pronunciation is concerned, the teachers generally emphasized how vital it was to pay attention to teaching pronunciation. For example, Teacher 6 noted on this point that: "It is very important to teach them [students] pronunciation; I focus on it a lot". However, to what extent teachers really focus on teaching pronunciation is questionable. Further in the interview, Teacher 6 indicated that pronunciation did not have as much attention as it would deserve because Level 2 and Level 3 students "don't know the basics, like pronunciation, they confuse words, they are not able to distinguish bed, bad and bat […], this implies that not sufficient emphasis is placed on teaching them to listen, to pronounce, in SLP1 courses". Interestingly, Teacher 6 was not the only one to admit this, with Teacher 5 saying: "I do exercises focusing on pronunciation. Some of my colleagues don't do that but I find it very important".

The teachers' talk concerning the development of discourse knowledge was mainly concerned with developing students' ability to listen for gist and a specific detail. The teachers generally emphasized the importance of playing the recording more than once as during the first hearing students "concentrate on general meaning" (Teacher 2) and during the second hearing they "focus on specific information" (Teacher 2).

4.3.1.2 Subtheme 2: development of strategic competence

As regards the cognitive strategies, the teachers' talk mainly focused on tasks fostering students' ability to predict general content of the text before listening. One example of this comes from Teacher 1, who, as a part of a pre-listening phase, teaches students how, "using the questions", they can "identify the content, who might talk and what their task is". Although inferencing was a cognitive strategy the teachers discussed most and not much or almost nothing was mentioned about the remaining strategies, judging from a relatively concrete description of procedures aimed at improving students' cognitive ability, it could perhaps be concluded that cognitive strategies are taught to students attending L1 courses.

Unlike with cognitive strategies, teachers' reflections on how they supported the development of their students' meta-cognitive strategies were rather vague. Although, when talking about an effective listening comprehension lesson, the great majority of the teachers mentioned pre-listening, while-listening and post-listening activities they used, it was not clear from the description they provided whether while performing the activities, students' meta-strategic awareness was developed. For example, when asked to specify methodological principles of teaching listening comprehension she followed, Teacher 2 mainly spoke about following "the teacher's book", where "there are some pre-listening tasks, then listening, followed by a certain kind of analysis". Teacher's 4 account of how she approached post-listening stage was mainly about checking the correctness of the responses: "We play the text from the beginning and where there is the correct response, I stop the recording and I always ask students whether they could hear it". With no remarks about whether students are provided a chance to evaluate their initial predictions, compare what they could understand and what not and talk about the strategies they could have used to arrive at a correct response, it would be unwise to conclude from the data that students receive effective meta-cognitive instruction.

The excerpts classified as being related to students' familiarity with the test format mostly reflected teachers' dissatisfaction with the amount of information these had about the test format. One of the teachers likened the test to the emperor's new clothes, "which everyone talks about but nobody has ever seen" (Teacher 7). This may come as a surprise for several reasons. First, the testing team of the LI regularly organizes seminars at which the teachers are made familiar with STANAG 6001 and the requirements for English language proficiency levels SLP1, SLP2 and SLP3. Second, the teachers have at their disposal a Level 1 exam handbook, which outlines basic structure of L1 listening test, presents an overview of the language knowledge and listening comprehension sub-skills required for Level 1 and provides samples of listening comprehension items. Third, there is Level 1 practice listening test, which reflects the actual test rather accurately.

With the exception of one teacher, who emphasised the need for more sample tests that would enable students to practise test-taking strategies, the teachers generally avoided the topic of developing test wiseness. Judging from the questionnaire results regarding test wiseness, it could be said that students are taught test taking strategies. The question, however, is why there was a significant decrease in the test-wiseness pre-test and post-test means (pre-test mean=4.88, post-test mean=4.50). Students' inability to apply test-taking strategies could have been one reason for the decrease. Another reason could have been too short pauses between the recordings on the test, which is something the teachers said students often complained about and is further discussed in Section 4.3.2.4.

4.3.2 Theme 2: constraints in teaching listening

During the interview, the teachers generally spent extensive time commenting on factors preventing them from teaching listening more efficiently. Remarks related to each other were grouped to five subthemes discussed in the subsequent sub-sections.

4.3.2.1 Subtheme 1: obstacles on the side of teachers

As regards the teachers, one of the main factors hindering the development of students' listening skills appears to be teachers' lack of theoretical background for teaching listening. This conclusion was deduced from a detailed examination of responses given to the questions aimed at finding out how the teachers actually teach listening. Most of their responses were too general, lacking specificity and detail one would probably expect. For example, when asked whether there were any methodological principles she adhered to when teaching listening, teacher 5 stated: "I think that there maybe are…With SLP1 students, I proceed very slowly and I maybe sometimes help them too much". Some of the teachers admitted their lack of relevant theoretical knowledge explicitly, by claiming that "nobody prepared us for it [teaching listening]. We are learning from experience and from what we study ourselves, which is natural. We are truly not prepared" (Teacher 1).

This seeming methodological uncertainty could maybe explain why most of the teachers rely on methodological techniques described in teachers' books, which are, according to Teacher 3, "methodologically very well designed". Although teacher's books undoubtedly provide teachers with great methodological support, they have been criticised on many occasions for failing to apply the results of scientific research focused on effective teaching methods of teaching English language (Dunlosky, 2013; Goh, 2008; Mendelsohn, 1998). Moreover, listening activities included in most of textbooks are topic-based. Thus, their main focus is not on developing students' listening strategies but on introducing, revising or fixing vocabulary or grammar related to the topic discussed (Jana Balíková, personal communication, July 20, 2016). It thus seems that it is up to a teacher alone to guide their students through the process of listening. It is worth mentioning here that despite the apparent over-reliance on textbooks, some of the teachers pointed to the importance of providing students with additional material through which they can practice listening. KET and PET test books were often mentioned in this regard.

With regard to selecting supplementary materials, some of the teachers pointed to problems they had with distinguishing between Level 1 and Level 2. Teacher 8 wondered whether “the teachers are not too demanding, pushing students somewhere where they get lost and students decide to pack it in”. The remark seems to suggest that high demands teachers place on students may be one of the main reasons why students are unmotivated and why there has been relatively low success rate for L1 listening test.

4.3.2.2 Subtheme 2: obstacles on the side of students

Unlike the student participants, who expressed strong motivation to study English in both versions of the questionnaire (pre-test median=6, post-test median=5), the teachers often complained about lack of student motivation. According to Teacher 6, it is very difficult to convince students that they can learn to listen. She stated that the main reason was that “unlike other skills, listening is completely out of their control - they cannot influence timing or the amount of time they have for writing the response – and that’s what scares them”. By saying that students “are motivated by the fact that they will take STANAG exam, so […] they have the reason to learn but they do not always do so because we are humans”, Teacher 1 suggested that although students feel motivated, their motivation comes from external factors and not from within, which could explain why “in the beginning students try hard, but later…they are terribly tired. The third month is a disaster […] they don’t want to do anything in the afternoons” (Teacher 8).

What may also have a negative impact on student motivation is the fact that, as argued by the teachers, the classes are heterogeneous, with some of the students being “true beginners” and others having “A-level in English” (Teacher 2). This places a burden on teachers as they have to adapt listening tasks to meet the needs of students with varying English language abilities.

Another fact that the teachers commonly agreed on in relation to the obstacles on the side of students was students’ low aptitude for learning, pushing

teachers to spend much of the time allotted to teaching listening on teaching their students how to actually learn.

4.3.2.3 Subtheme 3: obstacles on the side of the employer

There were two issues that the teachers discussed mostly with regard to this subtheme: short duration of the course and material constraints.

The teachers often argued that it was impossible to make “a person who has never learnt English comprehend English language in four months” (Teacher 7). Majority of the teachers claimed that the low success rate in L1 listening test stemmed from the short courses’ duration. The teachers’ opinion on L1 course’s duration was succinctly expressed by Teacher 5:

> Maybe it’s ironical but listening does not get as much attention as it does in SLP2 and SLP3 courses. It is because when students finally reach the level necessary for comprehending simple monologues or dialogues, we have little time left to practice.

Another hindrance to the development of students’ listening skills, about which the teachers often talked, was the material and technical conditions in which they worked. The teachers generally viewed their material and technical equipment as one of the main factors undercutting their efforts to use more effective and more modern methods of teaching listening. The following statement by Teacher 6 clearly illustrates the teachers’ disillusionment:

> The main problem is that all the activities and all the ideas are great, but unfeasible in our conditions because of the technical equipment. We are taught how to download materials from the net, to create our own activities but I am not able to do it using the net and the PC at work. I can do it at home, in my free time, on my net, on my personal PC […]. If I do so, problem No. 2 arises; I cannot play it in the classroom. So, it’s useless to design activities using materials downloaded from the net, as there are no devices in the classroom on which I could play them.

4.3.2.4 Subtheme 4: obstacles related to L1 listening test

Although most of the teachers thought the test results reflected students' true listening ability, all of them called for each recording included in the test to be played twice. They rationalized this opinion by saying that "there are too many activities for SLP1 students to do at the same time" (Teacher 4) as "they have to listen, to follow the text, they have to read, to write, they have a test and an answer sheet" (Teacher 5). The teachers believed that playing all recordings twice could ease students' mental overload and decrease their test anxiety. With respect to decreasing students' stress, the teachers also suggested prolonging the pauses between individual recordings on the test as well as providing students with extra time for reviewing their responses after completing the test.

Another issue that the teachers discussed concerning this subtheme was quality of the recordings. Teacher 2 remarked that her students often complained about the difference in the way the instructions and the recordings themselves were recorded, saying that: "The instructions were played loud enough, but when the test started the sound volume fell, but the test was in progress and there was nothing the students could do about it".

4.4 Classroom observation

Altogether 34 codes were developed during the thematic analysis of the observation data. These were then grouped into five main themes, all of which were related to the way students' listening ability was developed in the context of L1 courses. The main themes and their sub-themes are presented in Figure 5.

Although it was evident from the class observations that there was a clear connection between the content of the classes and the listening construct tested by L1 listening test, the results of the analysis revealed some deficiencies in teaching listening comprehension at Level 1. These included (a) little class time devoted to developing language competence, (b) shallow

strategy instruction, (c) lack of lesson planning, (d) problems choosing listening passages at the appropriate level of difficulty, and (e) teacher-centred lessons.

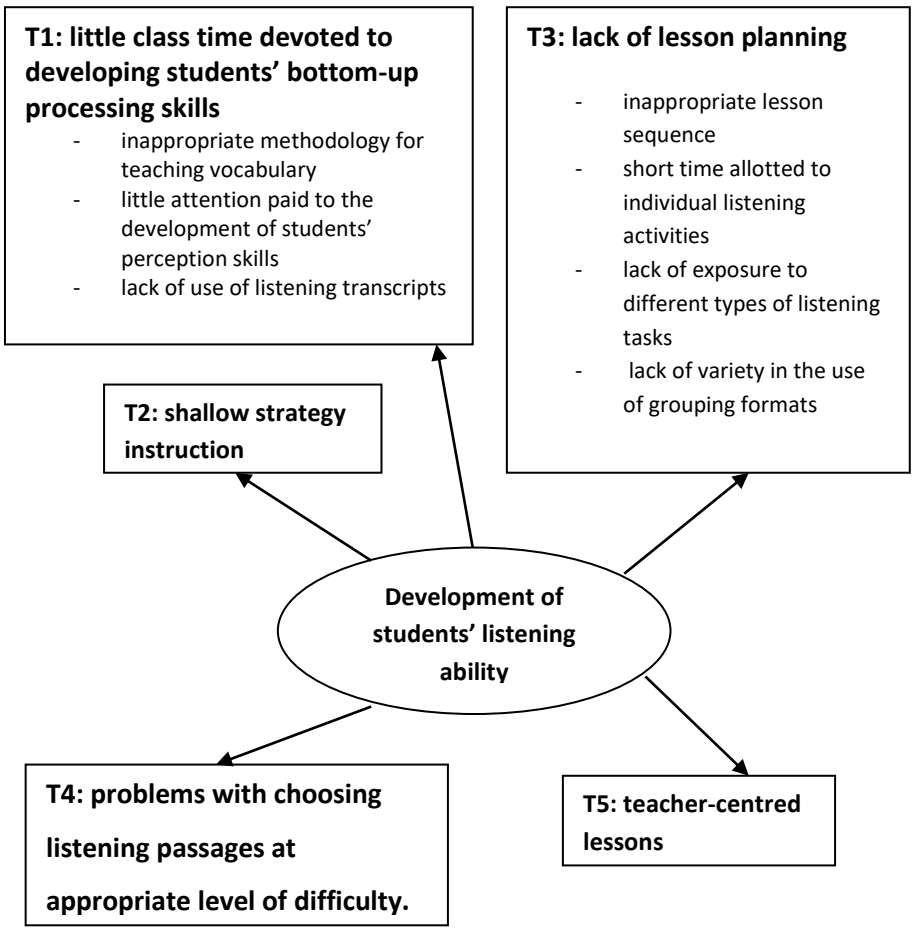


Figure 5. Themes and sub-themes that emerged from the thematic analysis of the observations

4.4.1 Theme 1: little class time devoted to developing the students' bottom-up processing skills

Perhaps the most notable fact about the classes observed was that a surprisingly considerable part of class time was devoted to trying to develop the students' meta-cognitive and cognitive strategies, which, considering the widely acknowledged positive impact of strategy instruction on L2/FL listening comprehension (e.g., Goh, 2008; Chen, 2013; Lynch & Mendelsohn, 2010; Vandergrift & Goh, 2009), is definitely something for which the teachers should be commended. As for meta-cognitive strategies, the focus was directed

mainly towards setting the purpose for listening, checking understanding of message while listening and self-evaluating. In case of the development of cognitive strategies, attention was mainly given to the development of top-down strategies, namely predicting general content before listening using the context, selective attention, inferencing, and summarizing.

However, relatively little proportion of class time was spent on enhancing the students' bottom-up processing skills, the acquisition and automatization of which is considered crucial to success in listening comprehension for L2/FL students (e.g., Goh, 2000; Vandergrift & Goh, 2009). This especially holds true for lower-proficiency students, whose comprehension of spoken language often breaks down on their inability to process oral input at lexical, syntactic or phonemic level (Field, 2003).

Despite the fact that, as for lexis, efforts were made by the teachers to develop their students' vocabulary, in the vast majority of the classes observed, translation was the only technique used for teaching vocabulary. The teachers rarely wrote the new words or phrases on the board and if they did, the phonetic transcription and the translation of the words into Slovak were not given. Also, with the exception of one case, the students were not given an opportunity to practise the new vocabulary and so there were little chances that the newly presented words and phrases got into their working vocabulary. This might have been caused by the fact that the observations were conducted relatively close to the test date, which seemed to have influenced the content of the classes, with their main focus being on developing test-taking strategies.

This could perhaps also explain why, given the relatively low proficiency of the students, too little attention was paid to helping the students develop their perception skills. Altogether, an attempt at developing the students' word-segmentation skills was observed only in two cases, which, taken into consideration the fact that the ability to discriminate sounds, recognize stress and intonation patterns and distinguish word boundaries is considered to be

one of the most important listening abilities for beginners (Brown, 2004), was far too little.

However, what appeared most surprising in this regard was that listening transcripts were distributed to the students on one occasion only. Although, when there was a disagreement among the students about the correct answers to individual questions, the teachers quoted the transcript to justify the correct answers, they did not distribute the transcripts to the students and thus actually failed to utilize the great potential of using transcripts in, for example, clarifying points of confusion or helping students to see the difference between the spoken and written forms of English words (Vandergrift & Goh, 2009).

4.4.2 Theme 2: shallow strategy instruction

As already mentioned in Section 4.4.1, there were real efforts made by the teachers to integrate strategy instruction into L1 courses. For example, one teacher introduced one of the listening activities by reviewing vocabulary central to the listening passage. Then the teacher told the students to read the comprehension questions and asked them to tell her what they thought the listening passage was going to be about and what their task would be. By doing so, the teacher set the purpose for listening. However, an explicit explanation of what listening strategies could be useful to accomplish the task at hand and why was not provided. In order to check the students' listening comprehension, the teacher played the recording again and then asked the students to tell her what they thought the correct answer to each of the multiple-choice comprehension questions was. She also asked for an explanation of why the particular option was correct and the other options were wrong. Nevertheless, the post-listening discussion was missing a critical aspect of strategy instruction, namely, discussion about the effectiveness of the strategies used (Gilakjani & Ahmadi, 2011; Vandergrift & Goh, 2009).

The other teachers approached teaching listening strategies in a similar spirit. It was often observed that while on some occasions, the teachers took time to promote their students' cognitive and meta-cognitive awareness by, for instance, asking the students to predict the content of the listening passages they would hear or to summarize the main points of the listening passage, on others, the teachers resorted to what Goh (2008) calls "comprehensive-based techniques" (p. 189) and they simply asked the students to listen to the recordings and checked the students' comprehension by having them answer comprehension questions.

4.4.3 Theme 3: lack of lesson planning

The important role of lesson planning in teaching L2/FL has been stressed by many scholars and researchers (e.g., Liyanage & Bartlett, 2010; Van Tuyen, 2015). The observations, however, revealed that the majority of the teachers (seven out of eight) delivered their classes without a lesson plan, which, taking into account the ability of a lesson plan to help teachers avoid making common teaching mistakes, is striking.

One of the most commonly observed teaching mistakes that were likely to stem from insufficient lesson planning was inappropriate lesson sequence. Although it was evident from the classes observed that the teachers had been thinking about how to structure the lesson, the fact that they did not put it down on paper might have caused that they sometimes had difficulty in sequencing individual listening activities, which often threw students into confusion. For example, as a pre-listening activity, one teacher let students read the constructed-response comprehension questions before they listened and then told them to guess what the missing words from the blanks might have been. After that, she asked the students a couple of warm-up questions related to the listening passage, which were meant to set the scene and activate their background schemata. Then she played the recording. However, since the students had meanwhile forgotten what their task was, the whole listening activity did not have its desired effect.

Another teaching mistake was relatively short time allotted to individual listening activities. At least in the half of the classes observed, the teachers rushed the students from one listening activity to another, without providing them with many chances to actually think about the listening process. Providing little time for the students to formulate their responses to questions, answering questions instead of the students, reading task instructions too fast for the level or providing far too little time for the students to read four-option multiple-choice questions were typical examples of this type of a teaching mistake.

Lack of exposure to different types of listening tasks and lack of variety in the use of grouping formats could perhaps also be attributed to poor lesson planning. Despite the fact that the teachers should be praised for their efforts to make use of different types of listening passages on a wide variety of Level 1 topics, across the classes observed, attention was predominantly paid to practising only two types of a listening task, namely listening for specific information and listening for important detail. Only scant attention was paid to practising listening for gist and listening for simple main idea, which are statistically the most demanding types of listening tasks for Level 1 students and their practice would, therefore, deserve more attention.

As regards the variety of grouping formats, with the exception of two teachers, the teachers employed whole-class instruction most of the time, which is not to imply that whole-class instruction does not facilitate learning. However, given the mixed-ability nature of the classes, the teachers should surely consider the use of small group instruction and “pairs” instructions more as they both allow differentiation in instruction and enable a teacher to tailor instruction to the needs of students having problems with listening comprehension (Santangelo & Tomlinson, 2015).

4.4.4 Theme 4: difficulty in choosing listening passages at the appropriate level of difficulty

The classroom observations seemed to have confirmed what the analysis of the questionnaires and the teacher interviews revealed about the teachers' ability to select level-appropriate materials for listening instruction. In most classes observed, a considerable proportion of materials used for listening instruction was above Level 1. In many cases, the students had evident problems comprehending the listening passages chosen as they were too fast, too long or contained lexicon that was above the targeted level of proficiency. This points to the need for providing the teachers with support in estimating the text's difficulty, for example, by means of organizing a practical seminar focusing on all the factors a person must take into account when determining whether a passage is at appropriate level of difficulty.

4.4.5 Theme 5: teacher-centred lessons

From the classroom observations, it was also clear that it was the teachers who were dominant and had a tight control of what happened in the classrooms. Most of the teachers gave great attention to mechanical practicing of basic listening sub-skills but very little attention to encouraging the students' active participation. Even when the students were involved in interaction, most of the time, it was teacher-student interaction, the main function of which was to answer the teachers' questions. Therefore, it is highly recommended that the teachers provide the students with more opportunities for real interaction and try to find a balance between teacher-centred and student-centred teaching.

5 Discussion of results

In chapter 5, the study's main findings are presented in light of the three research questions posed to explore the research topic. On the basis of the findings, conclusions are drawn about the link between L1 listening test and classroom practices of the teachers teaching L1 courses. The chapter also aims to offer recommendations about what could be done to improve the alignment between the way listening is tested and taught in the context of the AF SR. Before discussing the study's main findings and their possible implications, the chapter first outlines main limitations of the study.

First, although much effort was put into ensuring the study followed all the relevant research principles and practices as well as suggestions made by the supervisor, the fact that the study was conducted by a novice researcher might have influenced the quality of the research instruments design, the data collection and analysis as well as of the way the research findings were reported.

Another limitation, frequently discussed in relation to utilizing qualitative research methods (e.g., Dörnyei, 2007; Gass & Mackey, 2007; Lazarton, 2008), was the issue of researcher's subjectivity in collecting data and analyzing research results. In an attempt to minimize the influence of the researcher's beliefs on the research objectivity, four sources of data, both quantitative and qualitative, were used to examine the researched phenomenon.

The third limitation was generalisability. Since the study focused on exploring the link between teaching and testing listening in a specific learning context, neither its findings nor its conclusions could be generalised to other contexts. Despite its limitations, it is, however, still hoped that the findings could assist the parties involved in improving the quality of teaching and testing listening at Level 1 in the context of the AF SR.

The main intention of the study was to explore the alignment between the listening construct targeted by L1 listening test and the listening construct taught in L1 courses. Three research questions were posed to address the issue. The first research question sought to examine the extent to which L1 courses teach language knowledge as well as listening sub-skills and strategies measured by L1 listening test. For the purpose of arriving at the answer to this research question, the findings from the analysis of the data gathered from all the four data collection instruments used were grouped and analyzed based on individual facets of Buck's framework for describing listening ability (2001), including language competence, cognitive strategies and meta-cognitive strategies. The Buck's framework was used because, together with the descriptors for Level 1, it defines the construct targeted by L1 listening test. Although not being explicit elements of the construct measured, test format familiarity as well as test wiseness are believed by many to be closely related to successful performance (e.g., Bachman, 1990; Dunlosky, 2013; Gan, 2009), which was the main reason why the alignment between the abilities tested and the abilities taught was assessed also from the perspective of the students' test format familiarity and test wiseness.

The results of the analysis of the questionnaire data revealed that the aspect of listening ability the students felt least confident about was language competence (pre-test mean=3.92, post-test mean=4.02). This finding seems to be in line with the results yielded from the analysis of the teacher interviews and the observations, which pointed to a tension between the teachers' beliefs about teaching listening and their instructional practices. Although, in the teacher interviews, the teachers discussed the importance of developing their students' language competence and explained how they went about teaching lexis and pronunciation, the observations showed that, given the low proficiency of the students, there was too little class time devoted to enhancing the students' language competence. This could perhaps explain why some of the students expressed the need for developing listening skills gradually (10%) and adapting listening instruction to weaker students (4%) and why some of the teachers questioned the amount of attention the teachers pay to teaching pronunciation.

As regards the development of cognitive strategies, the observations found that relatively little class time was devoted to the development of the students' bottom-up processing skills, the possession of which is deemed vital for successful L2/FL listening comprehension (e.g., Gilakjani & Ahmadi, 2011; Goh, 2000; Vandergrift & Goh, 2009). In this regard, it seems important to reiterate that the fact that the observations were conducted close to the test date was highly likely to influence the way how the teachers delivered listening instruction and the observed classes may thus not have been representative of typical listening–focused classes. Nevertheless, the fact that, in the teacher interviews, the teachers' comments concerning the development of cognitive strategies were primarily focused on describing the development of top-down processing skills along with the fact that prior to taking L1 listening test, the students expressed only low satisfaction with the effectiveness of L1 courses as far as the development of their cognitive skills is concerned (pre-test mean=3.99) seem to confirm what the observations found about the way language competence and bottom-up processing skills are developed in L1 courses.

The qualitative analysis of the data gathered from the questionnaires, teacher interviews and observations showed that one of the possible reasons for the relative neglect of the development of language competence and bottom-up processes might have been short course duration. In the interviews, most of the teachers maintained that students' successful performance on L1 listening test stood or fell by course duration, with many of them claiming that it was impossible to make a true beginner understand English spoken word in four months. This leads to an assumption that it is most likely time pressure placed upon the teachers that make them rush through the curriculum, leaving too little time for the enhancement of students' basic listening micro-skills, such as sound discrimination or word boundaries identification. A further complication is added by the fact that the teachers seemed to place higher demands on the students than actually needed, introducing the students to more complex listening strategies although they still have trouble processing aural input.

To sum up, despite the fact that successful L2/FL listening comprehension is believed to be a process involving the deployment of both bottom-up and top-down strategies (e.g., Goh, 2008; Vandergrift 2007), for teachers teaching L1 courses it is important to have in mind that it is basic language knowledge and bottom-up processing that characterise language competence at Level 1, define the listening construct tested by L1 listening test and, not least importantly, constitute what less proficient L2/FL learners tend to rely on most when listening to their non-native language. The teachers teaching L1 courses should therefore provide more opportunities for students to develop their basic language knowledge and bottom-up processing skills to help them automate the sounds and words perception and, by doing so, freeing their mental resources for the employment of more complex listening strategies. This need for a more bottom-up oriented approach to teaching L2/FL listening is in agreement with a call for a greater attention to be given to bottom-up strategies when teaching listening expressed by some of the scholars (e.g., Field, 2003; Goh, 2008).

The last facet with the development of which the students expressed only weak satisfaction prior to taking the test was the level of their familiarity with the test format (pre-test mean=3.95). The students' views seemed to reflect the views of the teachers, who often complained about the lack of information they had about the test format. It is, however, worth mentioning here that, although not significant, there was an increase in the means of test format familiarity after the test (pre-test mean=3.95, post-test mean=4.31), indicating the student participants knew about the test more than they thought they had known. In addition, a significant increase was found between the pre-test and post-test means for the test difficulty scale (pre-test median=2; post-test median=3), suggesting the test was perceived by the student participants to be more difficult before than after taking it. The fact that L1 courses gave relatively ample attention to making the students familiar with the test format was also confirmed through observations, during which it was noticed that the materials and classroom activities used were all apparently chosen and adapted by the teachers to make the students familiar with the test format and to help them develop and practise their test-taking strategies. This seems to

support the findings by Green (2006) and Rashidi and Javanmardi (2011), who showed that test preparation courses tend to spend much time on test preparation.

Judging from the statistical analysis of the students' responses to the pre-test questionnaire, in which the students expressed moderate satisfaction with the development of their meta-cognitive strategies (pre-test mean = 4.77) and test wiseness (pre-test mean=4.88), it would seem that both meta-cognitive strategies and test wiseness are taught effectively in L1 courses. This should be undoubtedly considered a positive element of the listening instruction provided by L1 courses. Nevertheless, the thematic analysis of the teacher interviews and observations revealed that although there were elements of strategy instruction in the way the teachers taught listening, rather than grounded in theory, the incorporation of meta-cognitive and test-taking strategies into the classroom instruction seemed to be intuitive and rather inconsistent, which could perhaps also explain why there was a statistically significant decrease in the pre-test and post-test means of the meta-cognitive strategies scale (pre-test mean=4.77, post-test mean=4.35) and the test wiseness scale (pre-test mean=4.88, post-test mean=4.31), suggesting the students' difficulty in applying both meta-cognitive as well as test-taking strategies during the actual test.

A possible reason for the teachers' relatively unsystematic approach to teaching listening strategies could be lack of theoretical knowledge of teaching listening, which the teachers openly admitted in the teacher interviews and which was also revealed through the analysis of the observations. This suggests a clear need for seminars and workshops organized by the testing team of the LI for the teachers to be focused on teacher preparation for delivering strategy instruction, which would develop theoretical and practical knowledge and skills teachers need in order to integrate strategy instruction into listening instruction. Introducing the teachers to strategy instruction frameworks and providing them with practical tips and examples of how to teach listening strategies would be highly beneficial in this regard as it is

believed to provide teachers with all the tools necessary for designing and delivering efficient listening lessons (Mendelsohn, 2006).

The second research question was asked to explore if and how the experience of sitting the test changed the students' perception of the effectiveness of L1 courses in developing their language and strategic competence as far as listening is concerned. The results of the quantitative and qualitative questionnaire data analysis showed that the students were generally satisfied with the way L1 course developed their listening sub-skills and strategies. The pre-test and post-test means for individual sub-scales ranged from 3.92 to 4.88, suggesting weak to moderate satisfaction. The students confirmed their satisfaction with the effectiveness of L1 course in developing their listening comprehension in their responses to the questionnaire open-ended questions, where almost half of the participants (41%) expressed their appreciation of how listening was taught.

The experience of taking the test seemed to have no significant impact on the way the students perceived the overall effectiveness of L1 courses in developing their language competence, cognitive strategies and test format familiarity. However, in case of meta-cognitive scale and test-wiseness scale, a significant decrease in the means before and after sitting the test was found. One possible reason for the decrease could be the students' lack of ability to actually apply test-taking strategies, which may be attributed to three factors. First, listening strategies were found to be developed in a relatively unsystematic way, which was most likely caused by teachers' insufficient theoretical background for teaching listening, already discussed three paragraphs above. Second, in the vast majority of the classes observed, it was noted that the students were provided with only few opportunities to practise listening strategies, which was highly likely due to poor lesson planning discovered through the observations conducted. Third, hearing recordings only once, too short pauses between recordings and the alleged poor quality of recordings, the teachers claimed in the interviews that the students often complained about, could also explain the students' limited ability to apply listening strategies.

Practical implication resulting from the discussion above is clearly that besides the already discussed need for in-house teacher training courses aimed at making the teachers familiar with the fundamentals of strategy instruction, there also is an evident need for teacher training courses to show the teachers how to plan their lessons effectively in order to avoid teaching mistakes stemming from poor lesson planning. In addition, the study findings also pointed to the need for validation of L1 listening test, which could eventually lead to a more reliable and more valid assessment of FL listening comprehension in the context of the AF SR.

The third research question was aimed at investigating the relationship between the L1 course participants' perception of the effectiveness of L1 courses in developing their listening skills and their test performance. It was expected that there would be a positive correlation between the two variables. However, in contrast to the expectations, the results of the Spearman rank correlation revealed no relationship between the two variables, with the student participants being generally content with the way their listening ability was developed in L1 courses both before (90%) and after (87%) taking the test. A possible explanation for the insignificant result is that the students realized that successful listening comprehension in L2/FL depended on a number of aspects and that their failure on the test should therefore not have been blamed solely on L1 courses. An alternative explanation is that as the consequence of the former Communist regime, during which it was forbidden to question authorities, it is not in the nature of the Slovaks to criticise, let alone if it is a well-established educational system, in case of which there are evident efforts made by the teachers to improve language skills of students despite all the financial, material and technical obstacles they have to face.

6 Conclusion and implications for future research

According to some scholars and researchers in the field of applied linguistics (e.g., Alderson & Wall, 1993; Bailey, 1996), in order for a test to generate positive backwash on teaching and learning, it must meet several conditions, including authenticity, direct testing or wide and unpredictable sampling of the targeted construct (Hughes, 2003). A condition of major importance in this regard, however, seems to be a meaningful link between the abilities tested and the abilities taught.

This study attempted to explore the alignment between the construct of listening measured by L1 listening test and taught in L1 courses. A mixed methods approach, consisting of quantitative and qualitative analysis of data obtained from four different data collection instruments, was adopted to address the research issue. Based on the findings, it could overall be concluded that there seemed to be a clear alignment between the listening sub-skills and strategies taught in L1 courses and those measured by L1 listening test, with considerable efforts made by the teachers teaching L1 courses to direct the listening instruction towards the body of language as well as strategic knowledge targeted by the test. The study thus appears to be in line with former studies (e.g., Gan, 2009; Rashidi & Javanmardi, 2011; Yang & Badger, 2015) which indicated that although test preparation is an indisputable aspect of test preparation courses, it is not necessarily prioritized to the development of general aspects of L2/FL language ability. The study also seems to add further evidence to support the view held by some of the scholars (e.g., Alderson & Wall, 1993; Hughes, 2003; Popham, 2001) that a test may produce a beneficial backwash if there is a meaningful link between the abilities tested and the abilities taught.

The results of the present study, however, also pointed to several issues in how listening comprehension is taught and tested in the context of the AF SR, the removal of which would sure benefit all the stakeholders greatly. The most pressing issues related to teaching and testing listening at Level 1 revealed by the study include teachers' lack of theoretical knowledge for the teaching of

listening, lack of lesson planning, material and technical limitations and questioned quality of L1 listening test. It is believed that despite its limitations, the findings of the study might help trigger the implementation of changes, which could lead to the improvement of the system of teaching and testing listening at Level 1. It is first recommended that in-house teacher training courses aimed at promoting the development of skills teachers need for efficient listening strategy instruction and lesson planning be implemented. Another suggestion is that more financial resources should be allocated to technical equipment of the classrooms. Last but not least, Level 1 listening test should be validated.

Besides the above-mentioned implications for educational and testing practice, the study also seems to have several implications for future research. It is hoped that the study replication could help gain insights into the alignment between the way listening comprehension is taught and tested in other contexts, thus contributing to the allegedly neglected research area of (Gan, 2009; Rashidi & Javanmardi, 2011; Yang & Badger, 2015) the effectiveness of test preparation courses in promoting learners' actual language proficiency. Another issue worth investigating is the impact of teacher training courses aimed at developing teachers' skills in delivering effective listening strategy instruction on the way they teach listening. It would also be intriguing to examine how universities prepare their graduates for challenges involved in preparing L2/FL learners for taking a high-stake proficiency test as far as listening is concerned.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge, UK: Cambridge University Press.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129. http://dx.doi.org/10.1093/applin/14.2.115

Allen, I. E., & Seaman, C. A. (2007). Likert scales and data analyses. *Quality Progress, 7*. Retrieved June 7, 2016, from http://asq.org/quality-progress/2007/07/statistics/likert-scales-and-data-analyses.html

Allison, P. D. (2001). Missing data. *Sage University Papers Series on Quantitative Applications in the Social Sciences*, 7–136. Thousand Oaks, CA: Sage. Retrieved June 29, 2016, from http://people.oregonstate.edu/~acock/growth/NCFRgrowth/Allison,%20Missing%20Data%201.pdf

Anderson, J. R. (2009). *Cognitive psychology and its implication* (7th ed.). New York: Worth.

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

Bachman, L. F. (2004). Stating hypotheses and making statistical inferences. In L. F. Bachman (Ed.), *Statistical analysis for language assessment* (pp. 209–228). Cambridge, UK: Cambridge University Press. Retrieved January 23, 2016, from https://modules.lancaster.ac.uk/pluginfile.php/546397/mod_page/content/3/unit3/Bachman_Chapter_7.pdf

Bachman, L. F., & Palmer, S. F. (1996). *Language testing in practice.* Oxford, UK: Oxford University Press.

Bagheri, M., & Karami, S. (2014). The effect of explicit teaching of listening strategies and gender on EFL learners' IELTS performance*. Journal of Language Teaching and Research, 5*(6), 1387–1392. doi:10.4304/jltr.5.6.1387–1392

Bailey, K. M. (1996). Working for washback: a review of the washback concept in language testing. *Language Testing, 13*(3), 257–279. http://dx.doi.org/10.1177/026553229601300303

Baker, E. L. (2004). *Aligning curriculum, standards, and assessments: fulfilling the promise of school reform (CSE report 645).* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved August 10, 2016, from https://www.researchgate.net/publication/238669591_Aligning_Curriculum_Standards_and_Assessments_Fulfilling_the_Promise_of_School_Reform

Barker, L. (1971*). Listening behavior*. Englewood Cliffs, NJ: Prentice Hall.

Barta, E. (2010). Test takers' listening comprehension sub-skills and strategies. *Working Papers in Pedagogy*, *5*(4), 59–85. Retrieved March 29, 2016, from http://langped.elte.hu/WoPaLParticles/W4Barta.pdf

Bebler, A. (2010). Slovakia. In A. Bebler (Ed.), *NATO at 60: the post-cold war enlargement and the alliance's future* (pp. 45–47). Amsterdam: IOS Press BV.

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *Nursing Plus Open, 2,* 8–14. http://dx.doi.org/10.1016/j.npls.2016.01.001

Bowen, J. D., Madsen, H., & Hilferty, A. (1985). *TESOL techniques and procedures.* Boston, MA: Heinle & Heinle Publishers.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. http://dx.doi.org/10.1191/1478088706qp063oa

Brown, H. D. (2004). *Language assessment: principles and classroom practices*. New York: Longman.

Brown, J. D. (2011). Likert items and scales of measurement? *SHIKEN: JALT Testing & Evaluation SIG Newsletter, 15*(1), 10–14. Retrieved June 11, 2016, from http://jalt.org/test/PDF/Brown34.pdf

Buck, G. (1997). The testing of listening in a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education, Volume 7: language testing and assessment* (pp. 65–74). Dordrecht, NL: Kluwer Academic Publishers.

Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.

Caldwel, J. S. (2008). *Comprehension assessment: a classroom guide*. New York: Guilford Press.

Carr, N. T. (2011). *Designing and analysing language tests*. Oxford, UK: Oxford University Press.

Chamot, A. U. (1987). The learning strategies of ESL students. In A. Wenden & J. Rubin (Eds.), *Learning strategies in language learning* (pp. 71–83). Englewood Cliffs, NJ: Prentice-Hall.

Chen, A. H. (2013). EFL Listeners' strategy development and listening problems: a process-based study. *The Journal of Asia TEFL, 10*(3), 81–101. Retrieved March 26, 2016, from http://www.asiatefl.org/main/download_pdf.php?i=57&c=1404265349.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Lawrence Erlbaum Associates. Retrieved July 4, 2016, from http://lib.myilibrary.com.ezproxy.lancs.ac.uk/Open.aspx?id=493929

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). London, UK: Routledge-Falmer.

Creswell, J. W. (2014). *Educational research: planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Harlow, UK: Pearson.

Curtin, F., & Schultz, P. (1998). Multiple correlations and Bonferroni's correction. *Biological Psychiatry, 44*(8), 775–777. http://dx.doi.org/10.1016/S0006-3223(98)00043-2

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford, UK: Oxford University Press.

Dubeau. J. J, (2006*). Are we all on the same page?: an exploratory study of OPI ratings across NATO countries using the NATO STANAG 6001 scale.* Unpublished master thesis. Carleton University, Ottawa. Retrieved August 4, 2016, from http://www.natobilc.org/documents/Projects/OPIRaterSTUDY.pdf

Dunlosky, J. (2013). Strengthening the student toolbox: study strategies to boost learning. *American Educator*, *37*(3), 12–21. Retrieved March 23, 2016, from http://files.eric.ed.gov/fulltext/EJ1021069.pdf

Fahim, M., & Fakhri Alamdari, E. (2014). Maximizing learners' metacognitive awareness in listening through metacognitive instruction: an empirical study. *International Journal of Research Studies in Education*, *3*(3), 79–91. http://dx.doi.org/10.5861/ijrse.2014.762

Field, J. (2003). Promoting perception: lexical segmentation in second language listening. *ELT Journal, 57*(4), 325–334. http://dx.doi.org/10.1093/elt/57.4.325

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: research and practice in assessing second language* (pp. 77–151). Cambridge, UK: Cambridge University Press.

Gan, Z. (2009). IELTS Preparation course and student IELTS performance: a case study in Hong Kong, *Regional Language Centre Journal, 40*(1), 23–41. http://dx.doi.org/10.1177/0033688208101449

Gass, S. M., & Mackey, A. (2007). Survey-based research. In S. M. Gass & A. Mackey (Eds.), *Data elicitation for second and foreign language research* (pp. 148–163). Mahwah, NJ: Lawrence Erlbaum. Retrieved November 1, 2015, from http://lib.myilibrary.com.ezproxy.lancs.ac.uk/Open.aspx?id=391957

Gilakjani, A., & Ahmadi, A. (2011). A study of factors affecting EFL learners' English listening comprehension and the strategies for improvement. *Journal of Language Teaching and Research, 2*(5), 977–988. http://dx.doi.org/10.4304/jltr.2.5.977-988

Goh, C. M. A. (2000). A cognitive perspective on language learners' listening comprehension problems. *System, 28*, 55–75. Retrieved March 3, 2015, from https://modules.lancaster.ac.uk/pluginfile.php/420110/mod_page/content/10/listening-comprehension-problems.pdf

Goh, C. M. A. (2008). Metacognitive instruction for second language listening development: theory, practice and research implications. *RELC Journal, 39*(2). http://dx.doi.org/10.1177/0033688208092184

Green, A. (2006). Washback to the learner: learner and teacher perspectives on IELTS preparation course expectations and outcomes. *Assessing Writing: An International Journal, 11*(2), 113–34. http://dx.doi.org/10.1016/j.asw.2006.07.002

Green, A. (2007). Watching for washback: observing the influence of the international English language testing system academic writing test in the classroom. *Language Assessment Quarterly, 3*(4)*,* 333–368. http://dx.doi.org/10.1080/15434300701333152

Green, R. (2013). *Statistical analysis for language testers.* Basingstoke, UK: Palgrave Macmillan.

Hamilton, L. S. (2010). Testing what has been taught: helpful, high-quality assessments start with a strong curriculum. *American Educator*, *34*(4), 47–52. Retrieved March 23, 2016, from http://files.eric.ed.gov/fulltext/EJ909943.pdf

Handbook of English Language Proficiency Test Based on STANAG 6001, SLP 1 (n.d.). The Language Institute – General Staff of the Armed Forces of the Slovak Republic, Slovakia.

Harding, L., Alderson, J. C., & Brunfaut T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: elaborating on diagnostic principles. *Language Testing*, *32*(3), 317–336. http://dx.doi.org/10.1177/0265532214564505

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, UK: Cambridge University Press.

James, C. J. (1985). *The testing of proficiency in listening comprehension: guidelines for teaching.* Retrieved March 23, 2016, from http://files.eric.ed.gov/fulltext/ED262639.pdf

Lazaraton, A. (2008). Utilizing qualitative methods for assessment. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed.): *Vol. 7. Language testing and assessment* (pp. 225–239). New York: Springer Science+Business Media. Retrieved October 17, 2015, from http://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30424-3_176#page-1

Liyanage, I., & Bartlett, B. J. (2010). From autopsy to biopsy: a metacognitive view of lesson planning and teacher trainees in ELT. *Teaching and Teacher Education*, *26*, 1362–1371. http://dx.doi.org/10.1016/j.tate.2010.03.006

Lynch, T., & Mendelsohn, D. (2010). Listening. In N. Schmitt (Ed.), *An introduction to applied linguistics* (2nd ed., pp. 180–196). London, UK: Hodder Arnold.

Mendelsohn, D. (1998). Teaching the language skills. *Annual Review of Applied Linguistics, 18*, 81–101. http://dx.doi.org/10.1017/S0267190500003494

Mendelsohn, D. (2006). Learning how to listen using learning strategies. In E. Uso-Juan & A. Martinez-Flor (Eds.), *Current trends in the development and teaching of the four language skills* (pp. 75–90). Boston, MA: Heinle & Heinle. Retrieved July 31, 2016, from http://lib.myilibrary.com.ezproxy.lancs.ac.uk/Open.aspx?id=219413

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–256. http://dx.doi.org/10.1177/026553229601300302

Mellor-Clark, S., & Baker de Altamirano, Y. (2004). *Campaign English for the military level 1: student's book.* Oxford, UK: Palgrave Macmillan.

Murphy, R. (2007). *Essential grammar in use* (3rd ed.). Cambridge, UK: Cambridge University Press.

Nunan, D. (1997). Listening in language learning. *The Language Teacher*, *21*(9), 47–51. Retrieved, July 26, 2016, from http://jalt-publications.org/old_tlt/files/97/sep/nunan.html

O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition*. New York: Cambridge University Press.

Oxenden, C., Latham-Koenig, C., & Seligson, P. (2012). *New English file elementary: student's book.* Oxford, UK: Oxford University Press.

Pallant, J. (2007). *SPSS survival manual: a step by step guide to data analysis using SPSS for Windows* (3rd ed.). New York: Open University Press/McGraw-Hill Education.

Popham, W. J. (2001). Teaching to the test. *Educational Leadership, 58*(6), 16–20. Retrieved March 29, 2016, from http://olms.cte.jhu.edu/olms2/data/ck/file/TeachingtotheTest-Popham.pdf

Posner, D. (2004). What's wrong with teaching to the test? *Phi Delta Kappan, 85*(10), 749–751. http://dx.doi.org/10.1177/003172170408501009

Rashidi, N., & Javanmardi, F. (2011). The IELTS preparation washback on learning and teaching outcomes. *Cross-Cultural Communication*, *7*(3), 132–144. http://dx.doi.org/10.3968/j.ccc.1923670020110703.162

Riger, S., & Sigurvinsdottir, R. (2016). Thematic analysis. In L. A. Jason & D. S. Glenwick (Eds.), *Handbook of methodological approaches to community-based research: qualitative, quantitative and mixed methods* (pp. 33–41). Oxford, UK: Oxford University Press.

Rost, M. (2011). *Teaching and researching listening* (2nd ed.). London: Longman.

Rowntree, D. (1981). *Statistics without tears: an introduction for non-mathematicians.* Harmondsworth, UK: Penguin Books.

Rubin, J. (1994). A review of second language listening comprehension research. *The Modern Language Journal*, *78*(2), 199–221. http://dx.doi.org/10.1111/j.1540-4781.1994.tb02034.x

Santangelo, T., & Tomlinson, C. A. (2012). Teacher educators' perceptions and use of differentiated instruction practices: an exploratory investigation. *Action in Teacher Education, 34*(4), 309–327. http://dx.doi.org/10.1080/01626620.2012.717032

Solak, E. (2012). A study of needs analysis at Turkish Gendarmerie in terms of English for specific purposes. *The Journal of Language and Linguistic Studies*, *8*(2), 48–63. Retrieved August 8, 2016, from http://www.jlls.org/vol8no2/48-63.pdf

STANAG 6001 Edition 4 (2010). Retrieved August 8, 2016, from http://www.ji.mil.sk/data/files/3028.pdf

Subedi, B. P. (2016). Using Likert type data in social science research: confusion, issues and challenges. *International Journal of Contemporary Applied Sciences*, *3*(2), 36–49. Retrieved June 11, 2016, from http://ijcas.net/Files/CMSUserOwnFolder/issue/February-2016/02.pdf

Thompson, I. (1995). Testing listening comprehension. *AATSEEL Newsletter*, *37*(5), 24–31. Retrieved March 23, 2016, from http://files.eric.ed.gov/fulltext/ED386916.pdf

Trochim, M. K. (2002). *Research methods knowledge base*. Retrieved June 16, 2016, from http://anatomyfacts.com/Research/ResearchMethodsKnowledgeBase.pdf

Turner, C. E. (2014). Mixed methods research. In Kunnan, A. J. (Ed.), *The companion to language assessment* (pp. 1403–1417). Hoboken, NJ: Wiley-Blackwell. Retrieved October 17, 2015, from http://onlinelibrary.wiley.com.ezproxy.lancs.ac.uk/doi/10.1002/9781118411360.wbcla142/pdf

Turner, J. (2004). Language as academic purpose. *Journal of English for Academic Purposes, 3*(2)*,* 95–109. http://dx.doi.org/10.1016/S1475-1585(03)00054-7

Vandergrift, L. (1997). The strategies of second language (French) listeners: a descriptive study. *Foreign Language Annals*. 30(1), 387–409. http://dx.doi.org/10.1111/j.1944-9720.1997.tb02362.x

Vandergrift, L. (2007). Recent development in second and foreign language listening comprehension research. *Language Teaching, 40*(3), 191–210. http://dx.doi.org/10.1017/S0261444807004338

Vandergrift, L., & Goh, C. (2009). Teaching and testing listening comprehension. In M. Long & C. Doughty (Eds.), *The handbook of language teaching* (pp. 395–411). Oxford, UK: Blackwell. Retrieved March 26, 2016, from https://www.researchgate.net/publication/228007493_Teaching_and_Testing_Listening_Comprehension

Van Tuyen, L. (2015). An investigation of teaching effectiveness in tertiary English classes. *International Journal on Studies in English Language and Literature, 3*(4), 43–57. Retrieved August 14, 2016, from https://www.arcjournals.org/pdfs/ijsell/v3-i4/6.pdf

Wall, D. (n.d.). *Classroom observation*. Retrieved November 20, 2015, from https://modules.lancaster.ac.uk/mod/folder/view.php?id=350879

Wenden, A. L. (1991). *Learner strategies for learner autonomy*. Englewood Cliffs, NJ: Prentice-Hall.

Wolvin, A. D., & Coakley, C. G. (1985). *Listening* (2nd ed.). Dubuque, IA: William C. Brown Company.

Wu, Y. (1998). What do tests of listening comprehension test? – a retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15,* 21–44. Retrieved March 3, 2015, from https://modules.lancaster.ac.uk/pluginfile.php/420110/mod_page/content/10/tst_019%281%29.pdf

Yang, Y., & Badger, R. (2015). How IELTS preparation courses support students: IELTS and academic socialisation. *Journal of Further and Higher Education, 39*(4), 438–465. http://dx.doi.org/10.1080/0309877X.2014.953463

APPENDIX A

Statistics on test takers' performance on STANAG 6001 exam at Level 1 for the period from 2008 to 2014

Level of successfulness in SLP 1 tests

		LISTENING		SPEAKING		READING		WRITING	
		SLP 1		SLP 1		SLP 1		SLP 1	
Year	No. of candidates	No. of SC	%	No. of SC	%	No. of SC	%	No. of SC	%
2008	324	154	48	272	84	184	57	273	85
2009	122	63	52	102	84	69	57	105	87
2010	59	18	31	46	78	25	42	47	93
2011	53	23	43	45	83	29	55	41	77
2012	53	26	49	49	92	36	68	47	89
2013	93	57	63	79	87	84	89	80	88
2014	81	40	49	70	86	65	80	66	81
Total	**785**	**381**	***49***	**663**	***84***	**492**	***63***	**659**	***84***

SC = successful candidates

Note. Internal material of the English Methodology and Testing Branch of the Language Institute

APPENDIX B
Buck's framework for describing listening ability

Table 4.1. *A framework for describing listening ability*

Language competence: the knowledge about the language that the listener *brings to the listening situation. This will include both fully automated* procedural knowledge and controlled or conscious declarative knowledge. Language competence consists of:

- **Grammatical knowledge:** understanding short utterances on a literal semantic level. This includes phonology, stress, intonation, spoken vocabulary, spoken syntax.
- **Discourse knowledge:** understanding longer utterances or interactive discourse between two or more speakers. This includes knowledge of discourse features, such as cohesion, foregrounding, rhetorical schemata and story grammars, and knowledge of the structure of unplanned discourse.
- **Pragmatic knowledge:** understanding the function or the illocutionary force of an utterance or longer text, and interpreting the intended meaning in terms of that. This includes understanding whether utterances are intended *to convey ideas, manipulate, learn or are for creative expression,* as well as understanding indirect speech acts and pragmatic implications.
- ***Sociolinguistic knowledge:*** *understanding* the language of particular socio-cultural settings, and interpreting utterances in terms of the context of situation. This includes knowledge of appropriate linguistic forms and conventions characteristic of particular sociolinguistic groups, and the implications of their use, or non-use, such as slang, idiomatic expressions, dialects, cultural references, figures of speech, levels of formality and registers.

Strategic competence: includes the cognitive and metacognitive strategies, or executive processes, that fulfil the cognitive management function in listening. This is the ability to use language competence, and includes all the compensatory strategies used by second-language listeners. *It consists of:*

- **Cognitive strategies:** those mental activities related to comprehending and storing input in working memory or long-term memory for later retrieval;
  - **Comprehension processes:** associated with the processing of linguistic and non-linguistic input;
  - **Storing and memory processes**: associated with the storing of linguistic and non-linguistic input in working memory or long-term memory;
  - **Using and retrieval processes:** associated with accessing memory, to be readied for output;
- **Metacognitive strategies:** those conscious or unconscious mental activities that perform an executive function in the management of cognitive strategies;
  - **Assessing the situation:** taking stock of conditions surrounding a language task by assessing one's own knowledge, one's available internal and external resources and the constraints of the situation before engaging in the task;
  - **Monitoring:** determining the effectiveness of one's own or another's performance while engaged in a task;
  - **Self-evaluating:** determining the effectiveness of one's own or another's performance after engaging in the activity;
  - **Self-testing**: testing oneself to determine the effectiveness of one's own language use or the lack thereof.

Note. From „Assessing listening", by G. Buck, 2001, p. 104

APPENDIX C
Listening test specifications

LISTENING TEST

Purpose of the test

The aim of the Listening Test is to measure the overall listening comprehension ability of the candidate according to specific criteria and levels of understanding as designated by STANAG 6001 and interpreted by BILC.

Construct

The assessment aims at testing the language competency comprised of grammatical, discourse, pragmatic, and sociolinguistic knowledge.

Language competence: the knowledge about the language that the listener brings to the listening situation. This will include both fully automated procedural knowledge and controlled or conscious declarative knowledge.

Language competence consists of:

Grammatical knowledge: understanding short utterances on a literal semantic level. This includes phonology, stress, intonation, spoken vocabulary, spoken syntax.

Discourse Knowledge: understanding longer utterances or interactive discourse between two or more speakers. This includes the knowledge of discourse features, such as cohesion, foregrounding, rhetorical schemata and story grammars, and knowledge of the structure of unplanned discourse.

Pragmatic knowledge: the understanding the function or the illocutionary force of an utterance or longer text, and interpreting the intended meaning in terms of that. This includes understanding whether utterances are intended to convey ideas, manipulate, learn or are for creative expression, as well as understanding indirect speech acts and pragmatic implications.

Sociolinguistic knowledge: understanding the language of particular sociocultural settings, and interpreting utterances in terms of the context of situation. This includes knowledge of appropriate linguistic forms and conventions characteristic of particular sociolinguistic groups, and the implications of their use, or non-use, such as slang, idiomatic expressions, dialects, cultural references, figures of speech, levels of formality and registers.

(Buck (2001): Assessing listening, 104)

Sub-skills to be tested at base levels

Level 1 'Elementary'

- *Identifying topic*
- *Understanding main idea(s)*
- *Identifying specific information and details*

In addition to the above, the following sub-skills have been defined as operational in testing Level 2:

Level 2 'Fair' 'Limited working'

- *Understanding cohesion and coherence*

In addition to the above, the following sub-skills have been defined as operational in testing Level 3:

Level 3 'Good' 'Minimum professional'

- *Understanding implicit information*
- *Distinguishing the main idea from supporting detail(s)*
- *Understanding supporting opinion, hypothesising, clarification, argumentation*
- *Identifying attitudes, emotional overtones, and subtleties of speech*
- *Understanding humour*
- *Distinguishing between different stylistic levels?*

Text types

Level 1

- *Simple telephone messages*
- *Announcements*
- *Simple conversations, interviews*
- *Simple descriptions of people, places, and things*
- *Simple narrations*
- *Straightforward instructions and directions*

Level 2&3

- *Telephone calls*
- *Announcements*
- *Conversations, interviews*
- *Narrations and descriptions*

- *Media (news)*
- *Arrangements*
- *Briefings*
- *Lectures, talks, speeches*
- *Presentations*
- *Discussions, debates*

Texts at **Level 1** are unambiguous, simple and highly predictable with linear organisation and in a standard dialect. Speech on simple topics (food, lodging, transportation, shopping, family, interests, etc.) contains high frequency structural patterns and vocabulary. The speed of delivery is at the slower end of the normal speaking speed range. Some recordings may need to be re-recorded to slow down speech rate.

Texts at **Level 2** are defined as concrete and factual, referring to real-life situations and familiar topics. Speech is delivered at a normal rate with some repetition and rewording, in a standard dialect.

Texts at **Level 3** are defined as both concrete and abstract, referring to familiar and unfamiliar topics. Speech is both formal and informal, delivered with normal speed and clarity in a standard dialect.

Texts are a combination of authentic and semi-authentic texts, and recorded in accordance with copyright laws.
They vary in both types and content, comprising monologues, interactive discourse between two or more speakers, and different voices (male, female, old, young, etc.). Recordings contain variety of accents corresponding to standard variants of English native speaker accent, and to English non-native speaker accents that approximate to the norms of native speaker accents.

Topics of recordings

As according to BILC amplification, topics are **general** and **professional** ('professional' topics defined as non-discipline specific or 'General Military English')
….including (but not limited to):

- *Everyday survival and work-related topics*
- *Military and security issues*
- *Economic and political matters*
- *Scientific and technical issues*
- *Cultural and social issues*
- *Physical, political and economic geography*

The subject matter should not advantage or disadvantage certain groups of candidates (e.g. highly discipline specific topics) nor should it offend in areas as religion, politics or sex.

Task types

To be chosen from the following range as appropriate to the texts:

- *Multiple-choice questions (**4** options)*
- *Short answer questions (with responses restricted to no more than five words)*
- *Matching items, objects or attributes?*
- *Listing*
- *Sentence/Text/Table completion*
- *Labelling or completing diagrams/maps/pictures*

Each recording may have more than one task type.
Recordings are played once, some recordings, if necessary, may be played twice.

Distinguishing Level 1 & 2 & 3 text and item difficulty in Listening

This is a multifaceted process where the test team, on the basis of its experience, knowledge of audience and TLU of audience decides on the level of difficulty of this task/item.
The decision derives from the interaction of the following:

- *the type and topic of the text*
- *the sub-skills tackled*
- *complexity and frequency of the test questions*
- *length of text*
- *the organisation of text*
- *complexity and low/high frequency of vocabulary and structures*
- *density of ideas*
- *speed of delivery*
- *the number of speakers*
- *the amount of redundancy, paraphrasing*
- *register*

N.B. This can only be properly determined by **pre-testing** of the text and items.

Test/recording length and number of tasks/recordings

Minimum number of items per level is determined as 20.

The recordings selected for testing purposes will be of diverse length and difficulty to elicit candidate's language level. They contain pauses for either reading the texts or the task or for solving the task.

	Test B	Test A
Test level	0+ to 1	1+ to 3
Time	25 min /approx/	45 min /approx/
Number of items	Minimum 20	Minimum 40
Number of recordings	Up to 15	Up to 30
Number of items per recording	Up to 2	Up to 3
Number of voices	Up to 2 voices	Up to 4 voices
Length of recording	Up to 1,5 min	Up to 2,5 min

Rubrics

Instructions (including the examples) to be heard on the CD and the one to be read on the test are identical for the test. They are written and spoken in Slovak.

All steams and options (in case of MCQ) are given in English.

Candidates are given the information about the text type and source, and that the recording will be played twice (where applicable).

Criteria for marking

Each item carries one mark. No marks are given for partially correct answer. Linguistic accuracy or spelling will not be assessed provided errors do not hinder comprehension.

Test will be marked according to the marking scheme including all acceptable answers.

For the distribution of particular levels, see the chart below:

Level	**Test B** **/contains Level 1 items/**	**Test A** **/contains Level 2&3 items/**
0+	60 - 69% of all items	
1	**70 % of all items**	
1+		50 - 69% of Level 2 items
2		**min 70% of Level 2 items**
2+		80- 90% of Level 2 items & 60 - 69% of Level 3 items
3		**90 - 100% of Level 2 items & min 70% of Level 3 items**
3+		100% of items

The scale can be adjusted to compensate for any slight imbalance in levels of difficulty.

Description of typical performance at basic levels

LEVEL 0 (NO PROFICIENCY)

No practical understanding of the spoken language. Understanding is limited to occasional isolated words. No ability to comprehend communication.

LEVEL 1 (ELEMENTARY)

Can understand common familiar phrases and short simple sentences about everyday needs related to personal and survival areas such as minimum courtesy, travel, and workplace requirements when the communication situation is clear and supported by context. Can understand concrete utterances, simple questions and answers, and very simple conversations. Topics include basic needs such as meals, lodging, transportation, time, simple directions and instructions. Even native speakers used to speaking with non-natives must speak slowly and repeat or reword frequently. There are many misunderstandings of both the main idea and supporting facts. Can only understand spoken language from the media or among native speakers if content is completely unambiguous and predictable.

LEVEL 2 (LIMITED WORKING)

Sufficient comprehension to understand conversations on everyday social and routine job-related topics. Can reliably understand face-to-face speech in a standard dialect, delivered at a normal rate with some repetition and rewording, by a native speaker not used to speaking with non-natives. Can understand a wide variety of concrete topics, such as personal and family news, public matters of personal and general interest, and routine work matters presented through descriptions of persons, places, and things; and narration about current, past, and future events. Shows ability to follow essential points of discussion or speech on topics in his/her special professional field. May not recognise different stylistic levels, but recognises cohesive devices and organising signals for more complex speech. Can follow discourse at the paragraph level even when there is considerable factual detail. Only occasionally understands words and phrases of statements made in unfavourable conditions (for example, through loudspeakers outdoors or in a highly emotional situation). Can usually only comprehend the general meaning of spoken language from the media or among native speakers in situations requiring understanding of specialised or sophisticated language. Understands factual content. Able to understand facts but not subtleties of language surrounding the facts.

LEVEL 3 (MINIMUM PROFESSIONAL)

Able to understand most formal and informal speech on practical, social, and professional topics, including particular interests and special fields of competence. Demonstrates, through spoken interaction, the ability to effectively understand face-to-face speech delivered with normal speed and clarity in a standard dialect. Demonstrates clear understanding of language used at interactive meetings, briefings, and other forms of extended discourse, including unfamiliar subjects and situations. Can follow accurately the essentials of conversations among educated native speakers, lectures on general subjects and special fields of competence, reasonably clear telephone calls, and media broadcasts. Can readily understand language that includes such functions as hypothesising, supporting opinion, stating and defending policy, argumentation, objections, and various types of elaboration. Demonstrates understanding of abstract concepts in discussion of complex topics (which may include economics, culture, science, technology) as well as his/her professional field. Understands both explicit and implicit information in a spoken text. Can generally distinguish between different stylistic levels and often recognises humour, emotional overtones, and subtleties of speech. Rarely has to request repetition, paraphrase, or explanation. However, may not understand native speakers if they speak very rapidly or use slang, regionalisms, or dialect.

APPENDIX D
English translation of the pre-test version of the questionnaire

Pre-test questionnaire on the students' perception of their preparedness for passing a listening section of the STANAG SLP1 exam

Dear respondent, Code: _____

this questionnaire is a part of an MA dissertation, which explores the differences between testing and teaching listening at SLP 1. The main purpose of this questionnaire is to find out how you perceive your preparedness to successfully complete a listening section of STANAG SLP1 exam ***before*** *sitting it. Your opinion is important because the study findings will be used to improve methods and techniques of teaching and testing listening at the Language Institute. The questionnaire is divided into three parts and comprises 40 items. Its completion should not exceed 20 minutes. Please be assured that your answers will be kept anonymous and your identities will be protected. If you wish to get a brief summary of the findings, please contact me via e-mail: m.vargova@lancaster.ac.uk. Thank you for your time and cooperation.*

I. **This part contains a number of statements people might agree or disagree with. Next to each statement, please put in the box the number which best expresses the extent to which you agree with the statement. If, for example, you like cooking very much, put number „6" in the box.**

6-absolutely agree	**5-agree**	**4-partly agree**	**3-partly disagree**	**2-disagree**	**1-absolutely disagree**	
I like cooking.						6

Please answer each item on this questionnaire and choose only one option (the one that best describes your opinion).

6-absolutely agree	**5-agree**	**4-partly agree**	**3-partly disagree**	**2-disagree**	**1-absolutely disagree**	
1.	My knowledge of grammar is sufficient to pass the listening test.					
2.	My knowledge of vocabulary is sufficient to pass the listening test.					
3.	I know how many test items the test includes.					
4.	I know how many points I have to get to attain Level 1 in Listening.					
5.	We have been taught to read listening task instructions carefully when practising listening comprehension.					
6.	We have been taught how to use knowledge of the topic the recording is about to guess the meaning of unfamiliar words.					
7.	We have been taught how to use tone of voice or background sounds to guess the meaning of unfamiliar words in the recording.					
8.	When practising listening comprehension, we try to predict answers to the questions before playing the recording.					
9.	When practising listening comprehension, we try to identify the purpose of the listening task before playing the recording.					
10.	I will be able to comprehend a continuous flow of speech.					
11.	I will be able to respond to the test items aimed at identifying the main idea.					
12.	We have been taught to imagine a picture of what is heard when practising listening comprehension.					

6-absolutely agree	5-agree	4-partly agree	3-partly disagree	2-disagree	1-absolutely disagree	
13.	We have been taught how to use world knowledge to evaluate the logics of options when practising listening comprehension.					
14.	I know what type of tasks the test includes (e.g. short answer questions, constructed response tasks, multiple-choice tasks, table completion, etc.)					
15.	I know what language functions the test tests (e.g. listening for gist, listening for main idea, listening for important information, etc.)					
16.	We have been taught to identify key words in listening items when practising listening comprehension.					
17.	We have been taught to respond to test items even if we do not know the answer when practising listening comprehension.					
18.	When practising listening comprehension, we are instructed to check understanding by drawing on context.					
19.	When practising listening comprehension, we are instructed to check the correctness of our understanding against old and new information.					
20.	When practising listening comprehension, we have been taught to check and revise the answer immediately after answering each item.					
21.	We have been taught to avoid last minute changes when practising listening comprehension.					
22.	I will be able to respond to test items aimed at finding a specific detail (number, place, name).					
23.	I am able to guess the meaning of unknown words using the context.					
24.	When practising listening comprehension, we discuss about how we have arrived at the correct answer.					
25.	When practising listening comprehension, we discuss about how to overcome the problems occurring while listening.					
26.	We have been made familiar with the Listening Descriptors for Level 1.					
27.	We have been taught how to use world knowledge to evaluate the logics of options when practising listening comprehension.					
28.	We have been advised to translate some ideas from English to Slovak when practising listening comprehension.					
29.	I know what the format of recordings used in the test is (length, accent, monologue, dialogue, etc.)					
30.	I know what form the respond to the test items should take.					

II. **This part contains 10 questions. Please answer the questions by simply circling the word that best describes your feelings/opinion. In case of each question, please, specify reasons for choosing the option.**

How important is reading books to you?

Absolutely important	Important	Somewhat important	Somewhat Unimportant	Unimportant	Absolutely unimportant

Please specify: Because reading books is a great relax for me and I also learn a lot through reading.

31. How important is it for you to pass the exam?

Absolutely important	Important	Somewhat important	Somewhat unimportant	Unimportant	Absolutely unimportant

Please specify:_____

_____

_____

32. How motivated are you to study English?

Absolutely motivated	Motivated	Somewhat motivated	Somewhat unmotivated	Unmotivated	Absolutely unmotivated

Please specify:_____

_____

_____

33. How useful has STANAG Level 1 course been in developing your listening comprehension?

Absolutely useful	Useful	Somewhat useful	Somewhat useless	Useless	Absolutely useless

Please specify:_____

_____

_____

34. Was sufficient time devoted to practising listening during the course?

Yes	No

Please specify:_____

_____

_____

35. How difficult do you think the listening section of the test will be for you?

Very easy	Easy	Somewhat easy	Somewhat difficult	Difficult	Very difficult

Please specify:_____

_____

_____

36. Do you expect to pass the listening test?

Yes	No

Please specify:_____

_____

_____

37. Do you expect the test to be a fair measure of your listening skills?

Yes	No

Please specify: _____

_____

_____

38. Please state your suggestions on how teaching listening could be improved in future.

_____

_____

_____

39. Please state your suggestions on how testing listening could be improved in future.

_____

_____

_____

40. Please state any other comments you would like to make regarding teaching and testing listening.

_____

_____

_____

_____

III. **General Information**

What is your age? Please circle one of the options below.

18–25 years old	26–30 years old	31–35 years old	36–40 years old	41–45 years old	46–50 years old	over 50 years old	Prefer not to say

What is your gender? Please circle one of the options below.

Male	Female	Other	Prefer not to say

Did you learn English before you started attending STANAG SLP1 course? Please circle one of the options below.

Yes	No

If yes, please specify the duration of the period (in months or years):_____

_____

Were you deployed on a mission abroad before you started attending STANAG SLP1 course?

Yes	No

If yes, please specify the year of your deployment: _____

and the duration of your deployment (in months or years):_____

APPENDIX E
English translation of the post-test questionnaire

Pre-test questionnaire on students' perception of their preparedness for passing a listening section of STANAG SLP1 exam

Dear respondent, Code: _____

this questionnaire is a part of an MA dissertation, which explores the differences between testing and teaching listening at SLP 1. The main purpose of this questionnaire is to find out how you perceive your preparedness to successfully complete a listening section of STANAG SLP1 exam ***after*** *sitting it. Your opinion is important because the study findings will be used to improve methods and techniques of teaching and testing listening at the Language Institute. The questionnaire is divided into three parts and comprises 40 items. Its completion should not exceed 20 minutes. Please be assured that your answers will be kept anonymous and your identities will be protected. If you wish to get a brief summary of the findings, please contact me via e-mail: m.vargova@lancaster.ac.uk. Thank you for your time and cooperation.*

I. **This part contains a number of statements people might agree or disagree with. Next to each statement, please put in the box the number which best expresses the extent to which you agree with the statement. If, for example, you like cooking very much, put number „6" in the box.**

6-absolutely agree	**5-agree**	**4-partly agree**	**3-partly disagree**	**2-disagree**	**1-absolutely disagree**	
I like cooking.						6

Please answer each item on this questionnaire and choose only one option (the one that best describes your opinion).

6-absolutely agree	**5-agree**	**4-partly agree**	**3-partly disagree**	**2-disagree**	**1-absolutely disagree**	
1.	My knowledge of grammar was sufficient to pass the listening test.					
2.	My knowledge of vocabulary was sufficient to pass the listening test.					
3.	I had a good idea of how many test items the test includes.					
4.	I had a good idea of how many points I have to get to attain Level 1 in Listening.					
5.	Before each listening task, I carefully read the test task instructions.					
6.	I used knowledge of the topic the recordings were about to guess the meaning of unfamiliar words.					
7.	I used tone of voice or background sounds to guess the meaning of unfamiliar words in the recordings.					
8.	I tried to predict answers to the test items before the recording was played.					
9.	I tried to identify purpose of the test task before the recording was played.					
10.	I was able to recognize individual words in a flow of speech.					
11.	I was able to respond to the test items aimed at identifying the main idea.					
12.	Whilst listening to the recordings, I was imagining a picture of what I heard.					
13.	I used world knowledge to evaluate the logics of options.					
14.	I had a good idea of what type of tasks the test includes (e.g. short answer questions, constructed response tasks, multiple-choice tasks, etc.).					
15.	I had a good idea of what language functions the test tests (e.g. listening for gist, listening for main idea, listening for important information, etc.).					

6-absolutely agree	5-agree	4-partly agree	3-partly disagree	2-disagree	1-absolutely disagree	
16.	I tried to identify key words in the test items.					
17.	I responded to the test items even when I did not know the answer.					
18.	Whilst listening to the recordings, I checked my understanding by drawing on context.					
19.	Whilst listening to the recordings, I checked the correctness of my understanding against old and new information.					
20.	I checked and revised the correctness of my response immediately after answering each item.					
21.	I avoided last minute changes.					
22.	I was able to respond to the test items aimed at finding a specific detail (number, place, name).					
23.	I was able to guess the meaning of unknown words using the context.					
24.	After listening to the recording, I self checked the correctness of my listening comprehension.					
25.	After listening to the recording, I thought back to how I arrived at the response.					
26.	The knowledge of Listening descriptors for Level 1 helped me when taking the test.					
27.	I was able to remember the key points of the recordings.					
28.	I translated some ideas from English to Slovak whilst listening to the recordings.					
29.	I had a good idea of what format the recordings used in the test have (length, accent, monologue, dialogue, etc.).					
30.	I had a good idea of what form the response to the test items should take.					

II. **This part contains 10 questions. Please answer the questions by simply circling the word that best describes your feelings/opinion. In case of each question, please, specify reasons for choosing the option.**

How important is reading books to you?

Absolutely important	Important	Somewhat important	Somewhat Unimportant	Unimportant	Absolutely unimportant

Please specify: Because reading books is a great relax for me and I also learn a lot through reading._____
_____

31. How important is it for you to pass the exam?

Absolutely important	Important	Somewhat important	Somewhat unimportant	Unimportant	Absolutely unimportant

Please specify:_____
_____
_____

32. How motivated are you to study English?

Absolutely motivated	Motivated	Somewhat motivated	Somewhat unmotivated	Unmotivated	Absolutely unmotivated

Please specify: _____

_____

_____

33. How useful has STANAG Level 1 course been in developing your listening comprehension?

Absolutely useful	Useful	Somewhat useful	Somewhat useless	Useless	Absolutely useless

Please specify: _____

_____

_____

34. Was sufficient time devoted to practising listening during the course?

Yes	No

Please specify: _____

_____

_____

35. How difficult was the listening section of the test for you?

Very easy	Easy	Somewhat easy	Somewhat difficult	Difficult	Very difficult

Please specify: _____

_____

_____

36. Do you expect to pass the listening test?

Yes	No

Please specify: _____

_____

_____

37. Was the test a fair measure of your listening skills?

Yes	No

Please specify: _____

_____

_____

38. Please state your suggestions on how teaching listening could be improved in future.

_____

_____

_____

_____

39. Please state your suggestions on how testing listening could be improved in future.

_____

_____

_____

_____

40. Please state any other comments you would like to make regarding teaching and testing listening.

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Thank you for your willingness to complete the questionnaire. We believe that your opinions and suggestions will help to increase the quality of teaching and testing listening in the Language Institute.

APPENDIX F
Characteristics of the student participants

No. of cases	51	Frequency	Percent
Age (in years)			
	18-25	3	5.9
	26-30	6	11.8
	31-35	24	47.1
	36-40	16	31.4
	41-45	2	3.9
	Over 50	0	0.0
	Prefer not to say	0	0.0
Gender			
	Male	42	82.4
	Female	9	17.6
English language learning history			
	Yes	38	74.5
	No	13	25.5
How long learning English			
	less than a year	2	5.3
	about 1 year	1	2.6
	about 2 years	2	5.3
	about 3 years	9	24.7
	about 4 years	20	52.6
	over 4 years	4	10.5
When last learning English			
	less than a year ago	1	2.6

about 1 year ago	0	0.0
about 2 years ago	0	0.0
about 3 years ago	1	2.6
about 4 years ago	1	2.6
more than 4 years ago	35	92.1
Deployment abroad		
Yes	14	27,5
No	37	72,5

APPENDIX G
English translation of the semi-structured interview guideline

Research theme: *Listening Section of the NATO STANAG 6001 Level 1 Exam: How is the Tested Listening Construct Taught?*

Section I General introduction:

The main purpose of our interview today is to find out more about what you think about teaching and testing listening comprehension at Level 1. The interview is divided into three main parts: i) warm-up/factual questions; ii) questions aimed at your overall view on and experience with teaching and testing listening at Level 1; iii) a few final questions. The whole interview should not exceed 20 minutes.
Please, be assured that I am only interested in your opinion, so there are no "wrong" or "right" answers to any of the questions and you can feel free to be honest and critical. The interview will be recorded, but, please, be assured that all your answers will be kept confidential and your identity will remain anonymous.

Are there any questions before we start?

Section II Warm-up/factual Questions

1. How long have you been a foreign language teacher?
2. How long have you been working for the Language institute?
3. What level of courses do you teach (SLP1, SLP2, SLP3)?
4. How many SLP1 courses have you taught?

Section III Content questions

FIRST PRIORITY QUESTIONS		
Variable identified	**Question which could be used to address the variable**	**Issue/s that might arise**
1. *The teacher's usual way of teaching listening comprehension*	How do you usually go about teaching listening comprehension at Level 1? What is, in your opinion, an effective way of teaching listening comprehension?	Time constraints Physical constraints Lack of knowledge of listening theories
2. *Training in teaching listening comprehension*	Have you received any training in teaching listening? Would you welcome one? Why yes? Why not?	Lack of chance to get familiar with listening teaching methods and strategies Too theory-oriented training
3. *Challenges involved in teaching listening comprehension*	What are, in your opinion, the major challenges in teaching listening comprehension at Level 1?	Students' motivation Teachers' motivation (low salaries) Time constraints (high time pressure)

		Physical constraints Difference in students' English language proficiency
4. *Testing listening comprehension according to STANAG 6001*	What is your opinion on testing listening comprehension at Level 1 by STANAG 6001 Level 1 Listening Test?	Too high demands Most recordings played only once Low quality of the recordings used Fast rate of speech Lack of authenticity Lack of task-based assessment
5. Reasons for low success rate in SLP1 listening test	In past years, have you been satisfied with the success rate of your students on SLP1 listening tests? Why yes? Why not?	Difficult nature of listening Test anxiety Lack of automatization of a listening process Too high demands Lack of alignment between the listening construct tested and taught

Section IV Final questions

I have no more questions. Is there anything else you would like to add?

APPENDIX H
Semi-structured observation schedule (reduced to A4 size)

Listening Section of NATO STANAG 6001 Level 1 Exam: How is the Tested Listening Construct Taught?

OBSERVATION SCHEDULE

Date:		Objective of the lesson:	
Centre:		No. of students present:	
Teacher:		Lesson start:	
		Lesson end:	
		Teaching materials used:	

Activity	Start time/ Finish time	Skill focus	Sts' grouping format	Sts' level of interest	Notes	Post-Observation data analysis

Skill Focus: Listening (L), Reading (R), Writing (W), Speaking (S), Grammar (G), Vocabulary (V) **Sts' grouping format:** Individual (I), Pair (P), Group (G), Whole Class (WC) **Sts' level of interest:** High (H), Average (A), Low (L)

Listening Section of NATO STANAG 6001 Level 1 Exam: How is the Tested Listening Construct Taught?

Activity	Start time/ Finish time	Skill focus	Sts´ grouping format	Sts´ level of interest	Notes	Post-Observation data analysis

Skill Focus: Listening (L), Reading (R), Writing (W), Speaking (S), Grammar (G), Vocabulary (V) **Sts´grouping format:** Individual (I), Pair (P), Group (G), Whole Class (WC) **Sts´level of interest:** High (H), Average (A), Low (L)

APPENDIX I

English translation of one of the teacher interview transcripts

R: The main purpose of our interview today is to find out more about what you think about teaching and testing listening comprehension at Level 1. The interview is divided into three main parts: i) warm-up/factual questions; ii) questions aimed at your overall view on and experience with teaching and testing listening at Level 1; iii) a few final questions. The whole interview should not exceed 20 minutes. Please, be assured that I am only interested in your opinion, so there are no "wrong" or "right" answers to any of the questions and you can feel free to be honest and critical. The interview will be recorded, but, please, be assured that all your answers will be kept confidential and your identity will remain anonymous. Are there any questions before we start?

T: No.

R: How long have you been teaching English?

T: Since …., so it has been ….. years.

R: And how long have you been working at the Language Institute?

T: This has been the sixth year.

R: Which levels have you taught within this period?

T: All the three levels – SLP1, SLP2 and SLP3 – as well as specialised courses. However, I am most experienced in SLP2 level courses. I think I have taught at least four such courses. Then I have taught one Level 3 course and a few Level 1 courses.

R: How do you usually teach listening comprehension at Level 1?

T: Step by step. I begin with something students are familiar with, such as words that are similar in English and Slovak, too. We try to identify them in texts and textbooks. At first, students listen according to pictures. Then, before listening to more and more difficult texts, I try to explain what they are going to hear. We also try to identify what kind of text students are going to listen to according to instructions. I encourage them to guess who is going to speak and what they should do while listening. I usually try to teach them what part of speech they should expect, whether it is a noun, numeral or an adverb of place. In this way, students get more prepared for what to focus on while listening to the text.

R: So, this is how you prepare them for listening. Are there any other methods that you use?

T: Students usually listen to texts twice. At first, I want them to listen without making any notes, just focusing on the text. However, they do not always do so because some of them are too eager and they want to write down the answers as fast as possible. But we usually listen twice or even more times, especially when the text is harder and contains more difficult vocabulary or is recorded too fast. The methods that I use also depend on the kind of

listening tasks. The tasks sometimes require repeated listening. I sometimes try to use difficult texts as well. If the text is too demanding, I always prepare audio transcript that we work with so that students can understand what they should focus on to answer the tasks.

R: Are there any post-listening techniques that you use?

T: I try to find out what was difficult for the students. If it is possible, I try to play the part of the text again to identify what exactly was so hard. From my experience I can say that students understand the text and wonder what was so difficult after they have listened to the text several times.

R: How difficult it is to teach listening comprehension in comparison to other skills?

T: I think listening is the hardest because it depends on other than cognitive skills. It is about one's ability to hear. There are people who can hear well and people who cannot hear at all. Perception of what we hear is difficult even in our mother tongue because some people are not able to hear information. They just hear something and they cannot concentrate on what they should look for. And if they lack this skill in their mother tongue, it is difficult to develop it while learning English. Since obtaining listening skills depends on this ability, it is the hardest skill to learn in a foreign language.

R: Have you ever attended a course or a seminar focused on teaching listening skills?

T: I have attended only one such seminar that was organised by the English Methodology & Testing Branch of the LI. However, I do not remember having such seminars during my university studies. Nobody prepared us for it. We are learning from experience and from what we study ourselves, which is natural. We are truly not prepared.

R: Would you welcome a seminar focused on teaching listening comprehension?

T: I certainly would be grateful for a seminar with samples of listening tasks that would be difficult even for us. We need to learn how to work with such texts to understand how students feel when they face the text that contains new vocabulary and that is recorded faster. If I understand how students perceive the text, I will be able to explain the things so they can understand them.

R: What are the biggest challenges concerning teaching of listening skills at Level 1?

T: I often feel that students lose their motivation if they repeatedly fail in listening tasks. Motivation is a big challenge. We need to encourage students to listen to English also on their own and not to be disappointed if they do not understand what they hear. They need to try to understand the part of a song, basic information from news reports or anything else. Then they can see transcript of the news or lyrics of the song to verify if they understood properly. They really must listen on their own. Motivation is fundamental.

R: Are your students motivated?

T: They are motivated by knowing that they have to sit an exam. However, I do not think that they all are able to make themselves listen and learn. I can see some motivation but the

students' efforts do not always correspond to this motivation. They have the reason to learn, but they do not always do so. We are only human.

R: What is your opinion of testing listening skills at Level 1 according to STANAG 6001?

T: I am always and again surprised that students doing Level 1 test listen to texts just once. We know that some students who attend Level 1 courses passed the school-leaving examination from English in the past. However, we also have students who come as real beginners. Having learnt English for four months, they sit an exam completely stressed out. Apart from bad acoustics, they have to concentrate on three different things – they have to listen and simultaneously concentrate on listening tasks and the answer sheet. This is too much for them. They must focus on three things but they hear the text only one time. Before they are able to react and do what they are supposed to do, another text is played. As a result, they become increasingly stressed out because they realize that they have just failed to do the previous task. I suppose that in their mind they try to get back to what they have just done and if they have answered the task properly. Therefore, they fail to concentrate on another task. The problem is that texts are played too fast and they can hear them only one time. At Level 1 it is absolutely vital that students listen to texts two times. What is more, I believe that at the end of the test students should have at least 2 or 3 minutes to breathe out, to go over the twenty questions, to check their answers. They simply need some time to relax and to get the feeling that they have done everything they could to perform well on their exam.

R: Your suggestion is really good.

T: Students at Level 1 read the answers in the booklet and listen simultaneously. Even though it may be only three sentences, they fail to focus on what they should look for. They just have to do too many things at once, so they are not able to keep the information they need in their head and find the answers.

R: Are you satisfied with students' results at listening comprehension tests in recent years?

T: I have not taught L1 course for a long time, so I do not exactly know the test results. But when I taught these courses, the results were good. However, I have to say that students were quite sceptical after the exam because they thought they had failed it and that the exam was too difficult. In the end, however, the results were good.

R: Do you think that those results reflected their real knowledge?

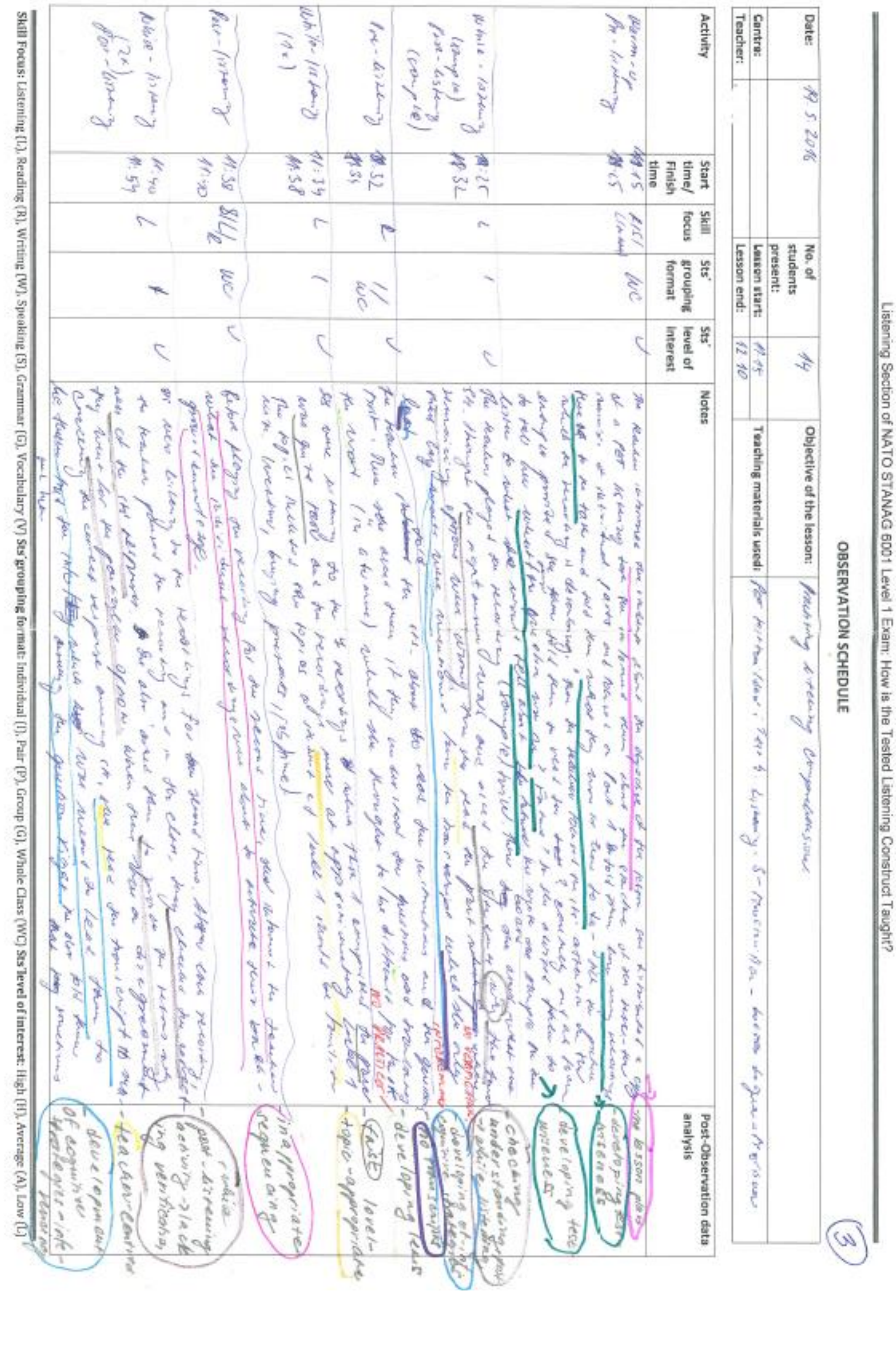
T: I think they did. After teaching one course for four months, we get to know our students quite well. I cannot assess other results, but I know that lately there have been some problems. Yet, I do not want to comment on them because I did not teach these courses.

R: We will see what the future holds. We keep our fingers crossed... Is there anything you would like to ask or say?

T: No.

R. Ok. Thank you for the interview.

APPENDIX J
Completed observation schedule


Listening Section of NATO STANAG 6001 Level 1 Exam: How is the Tested Listening Construct Taught?

OBSERVATION SCHEDULE

Date:		No. of students present:		Objective of the lesson:	
Centre:		Lesson start:		Teaching materials used:	
Teacher:		Lesson end:			

Activity	Start time/ Finish time	Skill focus	Sts' grouping format	Sts' level of interest	Notes	Post-Observation data analysis

Skill Focus: Listening (L), Reading (R), Writing (W), Speaking (S), Grammar (G), Vocabulary (V) Sts' grouping format: Individual (I), Pair (P), Group (G), Whole Class (WC) Sts' level of interest: High (H), Average (A), Low (L)

Activity	Start time/ Finish time	Skill focus	Sts' grouping format	Sts' level of interest	Notes	Post-Observation data analysis
Pre-listening activity (Task 2)	11:54/ 11:58	S/R/ L	WC/ I	✓	[illegible]	developing test-wiseness; no board use; no vocabulary practice; developing meta-cognitive skills: setting the purpose; fast pace; developing test format familiarity
While-listening activity (Task)	11:58 12:01	L	I	✓	[illegible]	fast, above level
Post-listening/ Pre-listening activity	12:02/ 12:02	L/C	WC	✓	[illegible]	
While-listening	12:03	L	I	✓	[illegible]	post-while-listening – lacking ...; no transcript; developing lexis; no board

Skill Focus: Listening (L), Reading (R), Writing (W), Speaking (S), Grammar (G), Vocabulary (V) Sts' grouping format: Individual (I), Pair (P), Group (G), Whole Class (WC) Sts' level of interest: High (H), Average (A), Low (L)

APPENDIX K

Mean, standard deviation, item-total correlation and Cronbach's alpha if item deleted for all the questionnaire items included in Section I

	Pre-test version – Cronbach's Alpha .87				Post-test version - Cronbach's Alpha .95			
Item No.	**M**	**SD**	**Item-Total Correlation**	**Cronbach's Alpha if Item Deleted**	**M**	**SD**	**Item-Total Correlation**	**Cronbach's Alpha if Item Deleted**
Language and strategic competence (Sub-scale 1)								
Cronbach's Alpha .77					**Cronbach's Alpha .90**			
1	4.14	.69	.71	.69	3.98	1.09	.81	.88
2	3.86	.78	.57	.71	3.84	.99	.83	.87
10	3.73	.78	.59	.70	4.08	1.00	.69	.89
11	3.78	.73	.52	.73	3.92	.87	.78	.88
22	4.26	.82	.14	.82	4.37	1.18	.60	.91
23	3.82	.93	.363	.71	3.75	.98	.63	.88
Cognitive strategies (Sub-scale 2)								
Cronbach's Alpha .70					**Cronbach's Alpha .80**			
6	4.84	.97	.39	.67	4.71	1.00	.61	.76
7	3.88	.82	.51	.64	3.55	1.05	.43	.79
12	4.26	1.01	.36	.68	4.31	1.03	.68	.74
13	4.12	.89	.49	.64	4.14	1.00	.73	.73
27	3.73	1.10	.47	.64	3.84	1.12	.62	.75
28	3.40	1.15	.38	.67	3.86	1.33	.34	.83
Meta-cognitive strategies (Sub-scale 3)								
Cronbach's Alpha .82					**Cronbach's Alpha .66**			
8	4.18	1.45	.49	.83	3.82	1.29	-.12	.81
9	4.69	1.16	.65	.78	4.82	.89	.58	.56

Item No.	M	SD	Item-Total Correlation	Cronbach's Alpha if Item Deleted	M	SD	Item-Total Correlation	Cronbach's Alpha if Item Deleted
18	4.78	.88	.76	.76	4.45	.99	.57	.56
19	4.51	.95	.69	.78	4.26	1.18	.49	.58
24	5.10	.85	.64	.79	4.02	1.12	.52	.57
25	4.65	1.02	.46	.82	4.08	1.04	.55	.56
Test format familiarity (Sub-scale 4)								
Cronbach's Alpha .80					**Cronbach's Alpha .87**			
3	3.08	1.85	.45	.80	4.04	1.53	.61	.87
4	3.78	1.83	.59	.76	4.18	1.30	.63	.86
14	4.69	1.27	.61	.76	4.45	1.19	.74	.84
15	4.59	1.27	.68	.74	4.47	1.08	.80	.83
29	2.96	1.17	.43	.79	4.16	1.03	.60	.86
30	3.81	1.31	.66	.74	4.22	1.05	.75	.84
Test-wiseness (Sub-scale 5)								
Cronbach's Alpha .74					**Cronbach's Alpha .77**			
5	5.73	.70	.46	.71	5.10	.88	.39	.76
16	5.24	.76	.55	.69	4.67	1.09	.60	.71
17	5.24	.91	.59	.67	4.80	1.20	.62	.71
20	4.45	1.10	.52	.69	4.39	1.08	.67	.70
21	3.96	1.13	.51	.70	4.04	1.33	.43	.77
26	4.61	.85	.29	.75	3.77	.99	.41	.76

APPENDIX L
Section I - skewness and kurtosis tests for normality

	Pre-test version	Post-test version
	Language and strategic competence	
Skewness	-.551	.014
Std. Error of skewness	.347	.347
Skewness/Std. Error of skewness	**1.59**	**.04**
Kurtosis	.664	-.365
Std. Error of Kurtosis	.681	.681
Kurtosis/Std. Error of Kurtosis	**.98**	**-.54**
	Cognitive strategies (Sub-scale 2)	
Skewness	.121	.107
Std. Error of skewness	.347	.347
Skewness/Std. Error of skewness	**.35**	**.31**
Kurtosis	-.571	.054
Std. Error of Kurtosis	.681	.681
Kurtosis/Std. Error of Kurtosis	**.84**	**.08**
	Meta-cognitive strategies (Sub-scale 3)	
Skewness	-.022	-.196
Std. Error of skewness	.347	.347
Skewness/Std. Error of skewness	**-.06**	**-.56**
Kurtosis	-.949	-.407
Std. Error of Kurtosis	.681	.681
Kurtosis/Std. Error of Kurtosis	**-1.39**	**-.60**

	Test format familiarity (Sub-scale 4)	
Skewness	-.078	-.306
Std. Error of skewness	.347	.347
Skewness/Std. Error of skewness	**-.23**	**-.88**
Kurtosis	-.158	-.609
Std. Error of Kurtosis	.681	.681
Kurtosis/Std. Error of Kurtosis	**-.23**	**-.89**
	Test-wiseness (Sub-scale 5)	
Skewness	-.128	-.118
Std. Error of skewness	.347	.347
Skewness/Std. Error of skewness	**-.37**	**-.34**
Kurtosis	-.136	-.266
Std. Error of Kurtosis	.681	.681
Kurtosis/Std. Error of Kurtosis	**-.20**	**-.39**

APPENDIX M
Section I - Kolmogorov-Smirnov test

Tests of Normality – Pre-test

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Total_LanguageCompetence	.115	47	.151	.963	47	.137
Total_CognitiveStrategies	.132	47	.040	.958	47	.092
Total_MetaCognitiveStrategies	.123	47	.072	.956	47	.072
Total_TestFormatFamiliarity	.124	47	.066	.981	47	.651
Total_TestWiseness	.096	47	.200*	.984	47	.760

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tests of Normality – Post-test

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Total_LanguageCompetence	.069	47	.200*	.989	47	.933
Total_CognitiveStrategies	.111	47	.187	.976	47	.443
Total_MetaCognitiveStrategies	.109	47	.200*	.979	47	.546
Total_TestFormatFamiliarity	.117	47	.122	.968	47	.224
Total_TestWiseness	.137	47	.028	.969	47	.247

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

APPENDIX N
Section I – results of the paired-samples t-tests

Language competence

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	PreTest_Total_LanguageCompetence	23.4894	47	3.28946	.47982
	PostTest_Total_LanguageComptence	24.1277	47	4.42138	.64492

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	PreTest_Total_LanguageCompetence & PostTest_Total_LanguageComptence	47	.011	.944

Paired Samples Test

		Paired Differences			
					95% Confidence Interval of the Difference
		Mean	Std. Deviation	Std. Error Mean	Lower
Pair 1	PreTest_Total_LanguageCompetence - PostTest_Total_LanguageComptence	-.63830	5.48288	.79976	-2.24813

Paired Samples Test

		Paired Differences			
		95% Confidence Interval of the Difference			
		Upper	t	df	Sig. (2-tailed)
Pair 1	PreTest_Total_LanguageCompetence - PostTest_Total_LanguageComptence	.97154	-.798	46	.429

Cognitive strategies

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	PreTest_Total_CognitiveStrategies	23.9149	47	3.69976	.53967
	PostTest_Total_CognitiveStrategies	24.6596	47	3.77219	.55023

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	PreTest_Total_CognitiveStrategies & PostTest_Total_CognitiveStrategies	47	-.111	.457

Paired Samples Test

		Paired Differences				
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference	
					Lower	
Pair 1	PreTest_Total_CognitiveStrategies - PostTest_Total_CognitiveStrategies	-.74468	5.56959	.81241	-2.37997	

Paired Samples Test

		Paired Differences				
		95% Confidence Interval of the Difference				
		Upper	t	df	Sig. (2-tailed)	
Pair 1	PreTest_Total_CognitiveStrategies - PostTest_Total_CognitiveStrategies	.89061	-.917	46	.364	

Meta-cognitive strategies

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	PreTest_Total_MetaCognitiveStrategies	23.8298	47	3.60144	.52532
	PostTest_Total_MetaCognitiveStrategies	21.7660	47	3.24515	.47335

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	PreTest_Total_MetaCognitiveStrategies & PostTest_Total_MetaCogniti veStrategies	47	.006	.969

Paired Samples Test

		Paired Differences			
					95% Confidence Interval of the Difference
		Mean	Std. Deviation	Std. Error Mean	Lower
Pair 1	PreTest_Total_MetaCognitiveStrategies - PostTest_Total_MetaCognitiveStrategies	2.06383	4.83378	.70508	.64458

Paired Samples Test

		Paired Differences				
		95% Confidence Interval of the Difference				
		Upper	t	df	Sig. (2-tailed)
Pair 1	PreTest_Total_MetaCognitiveStrategies - PostTest_Total_MetaCognitive Strategies	3.48308	2.927	46	.005

Test format familiarity

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	PreTest_Total_TestFormatFamiliarity	23.6809	47	5.43382	.79260
	PostTest_Total_TestFormatFamiliarity	25.8723	47	4.84372	.70653

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	PreTest_Total_TestFormatFamilirity & PostTest_Total_TestFormatFamiliarity	47	-.069	.643

Paired Samples Test

		Paired Differences				
					95% Confidence Interval of the Difference	
		Mean	Std. Deviation	Std. Error Mean	Lower	
Pair 1	PreTest_Total_TestFormatFamiliarity & PostTest_Total_TestFormatFamiliarity	-2.19149	7.52572	1.09774	-4.40112	

Paired Samples Test

		Paired Differences			
		95% Confidence Interval of the Difference			
		Upper	t	df	Sig. (2-tailed)
Pair 1	PreTest_Total_TestFormatFamiliarity - PostTest_Total_TestFormatFamiliarity	.01814	-1.996	46	.052

Test-wiseness

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	PreTest_Total_TestWiseness	29.2766	47	3.34080	.48731
	PostTest_Total_TestWiseness	26.3617	47	3.75572	.54783

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	PreTest_Total_TestWiseness & PostTest_Total_TestWiseness	47	.167	.262

Paired Samples Test

		Paired Differences			
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference
					Lower
Pair 1	PreTest_Total_TestWiseness - PostTest_Total_TestWiseness	2.91489	4.59125	.66970	1.56685

Paired Samples Test

		Paired Differences			
		95% Confidence Interval of the Difference			
		Upper	t	df	Sig. (2-tailed)
Pair 1	PreTest_Total_TestWiseness - PostTest_Total_TestWiseness	4.26293	4.353	46	.000

APPENDIX O
Section II - skewness and kurtosis tests for normality

	Pre-test version	Post-test version
	Question 31	
Skewness	-1.230	-2.483
Std. Error of skewness	0.333	0.333
Skewness/Std. Error of skewness	**-3.694**	**-7.456**
Kurtosis	1.290	9.625
Std. Error of Kurtosis	0.656	0.656
Kurtosis/Std. Error of Kurtosis	**1.966**	**14.672**
	Question 32	
Skewness	-0.591	-2.727
Std. Error of skewness	0.333	0.333
Skewness/Std. Error of skewness	**-1.775**	**-8.189**
Kurtosis	0.918	10.921
Std. Error of Kurtosis	0.656	0.656
Kurtosis/Std. Error of Kurtosis	**1.399**	**16.647**
	Question 33	
Skewness	-1.176	-1.509
Std. Error of skewness	0.333	0.337
Skewness/Std. Error of skewness	**-3.531**	**-4.478**
Kurtosis	5.092	3.024
Std. Error of Kurtosis	0.656	0.662
Kurtosis/Std. Error of Kurtosis	**7.762**	**4.568**
	Question 34	
Skewness	0.767	-0.050

Std. Error of skewness	0.333	0.333
Skewness/Std. Error of skewness	**2.303**	**-0.150**
Kurtosis	0.209	-0.204
Std. Error of Kurtosis	0.656	0.656
Kurtosis/Std. Error of Kurtosis	**0.318**	**-0.311**
	Question 35	
Skewness	0.201	-0.353
Std. Error of skewness	0.333	0.333
Skewness/Std. Error of skewness	**0.604**	**-1.060**
Kurtosis	-0.754	-0.531
Std. Error of Kurtosis	0.656	0.656
Kurtosis/Std. Error of Kurtosis	**-1.149**	**-0.809**
	Question 36	
Skewness	-0.102	-0.356
Std. Error of skewness	0.333	0.333
Skewness/Std. Error of skewness	**-0.306**	**-1.069**
Kurtosis	-0.503	-0.633
Std. Error of Kurtosis	0.656	0.656
Kurtosis/Std. Error of Kurtosis	**-0.767**	**-0.965**
	Question 37	
Skewness	-0.512	-0.493
Std. Error of skewness	0.333	0.337
Skewness/Std. Error of skewness	**-1.537**	**-1.463**
Kurtosis	-0.954	-0.659
Std. Error of Kurtosis	0.656	0.662
Kurtosis/Std. Error of Kurtosis	**-1.454**	**-0.995**

APPENDIX P
Section II - Kolmogorov-Smirnov test

Tests of Normality – Pre-test

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
q31_motivation	.302	51	.000	.750	51	.000
q32_motivation	.340	51	.000	.797	51	.000
q33_totalEffectivenes	.292	50	.000	.851	50	.000
q34_totalEffectiveness	.471	50	.000	.530	50	.000
q35_perceivedProspectOfSuccess	.228	51	.000	.878	51	.000
q36_perceivedProspectOfSuccess	.403	48	.000	.614	48	.000
q37_perceivedTestFairness	.370	50	.000	.632	50	.000

a. Lilliefors Significance Correction

Tests of Normality – Post-test

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
q31_motivation	.250	50	.000	.793	50	.000
q32_motivation	.372	50	.000	.726	50	.000
q33_totalEffectivenes	.217	47	.000	.895	47	.000
q34_totalEffectiveness	.439	46	.000	.579	46	.000
q35_perceivedProspectOfSuccess	.245	51	.000	.866	51	.000
q36_perceivedProspectOfSuccess	.360	48	.000	.634	48	.000
q37_perceivedTestFairness	.344	45	.000	.637	45	.000

a. Lilliefors Significance Correction

APPENDIX R
L1 listening test reliability analysis

Case Processing Summary

		N	%
Cases	Valid	51	100.0
	Excluded[a]	0	.0
	Total	51	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
.859	20

Item Statistics

	Mean	Std. Deviation	N
q1	.667	.4761	51
q2	.353	.4826	51
q3	.882	.3254	51
q4	.784	.4154	51
q5	.549	.5025	51
q6	.882	.3254	51
q7	.471	.5041	51
q8	.392	.4931	51
q9	.961	.1960	51
q10	.549	.5025	51
q11	.863	.3475	51
q12	.314	.4686	51
q13	.745	.4401	51
q14	.843	.3673	51
q15	.647	.4826	51
q16	.725	.4507	51
q17	.373	.4883	51
q18	.667	.4761	51
q19	.745	.4401	51
q20	.569	.5002	51

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
q1	12.314	18.900	.496	.850
q2	12.627	19.558	.325	.858
q3	12.098	19.370	.595	.849
q4	12.196	18.801	.612	.846
q5	12.431	19.690	.277	.860
q6	12.098	19.930	.394	.855
q7	12.510	19.095	.416	.854
q8	12.588	19.327	.371	.856
q9	12.020	20.420	.407	.856
q10	12.431	19.610	.296	.859
q11	12.118	19.586	.479	.852
q12	12.667	19.507	.351	.856
q13	12.235	19.224	.457	.852
q14	12.137	19.121	.599	.848
q15	12.333	18.387	.618	.845
q16	12.255	18.074	.757	.840
q17	12.608	18.723	.525	.849
q18	12.314	18.740	.537	.849
q19	12.235	19.584	.360	.856
q20	12.412	19.487	.327	.858

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
12.980	21.180	4.6021	20