DOKTORI (PhD) DISSZERTÁCIÓ

KÖVÉR ÁRMIN

“ALL EYES AND EARS”: INVESTIGATING FOREIGN LANGUAGE USERS’ PERFORMANCE IN LISTENING COMPREHENSION AND AUDIO-VISUAL COMPREHENSION

„CSUPA SZEM ÉS FÜL VAGYOK”: A HALLOTT SZÖVEGÉRTÉS ÉS AZ AUDIOVIZUÁLIS SZÖVEGÉRTÉS VIZSGÁLATA AZ IDEGENNYELV-HASZNÁLÓI PERFORMANCIÁK TÜKRÉBEN

2020

Eötvös Loránd Tudományegyetem Pedagógiai és Pszichológiai Kar

Neveléstudományi Doktori Iskola

Nyelvpedagógia Doktori Program

Vezetője: Prof. Dr. Károly Krisztina DSc, egyetemi tanár

"All Eyes and Ears": Investigating Foreign Language Users' Performance in Listening Comprehension and Audio-visual Comprehension

„Csupa szem és fül vagyok": A hallott szövegértés és az audiovizuális szövegértés vizsgálata az idegennyelv-használói performanciák tükrében

DOKTORI (PhD) DISSZERTÁCIÓ

Kövér Ármin

Témavezető: Dr. Dávid Gergely, habilitált egyetemi docens, ELTE BTK

A bíráló bizottság elnöke:	Prof. Dr. Medgyes Péter, professor emeritus, ELTE BTK
Belső opponens:	Dr. Brózik-Piniel Katalin, egyetemi adjunktus, ELTE BTK
Külső opponens:	Dr. Szabó Gábor, habil. egyetemi docens, PTE BTK
A bizottság titkára:	Dr. Veljanovszki Dávid, egyetemi adjunktus, ELTE BTK
A bizottság további tagjai:	Dr. Kolláth Katalin, főiskolai tanár, BGE
	Dr. Király Zsolt, habil. egyetemi docens, ELTE BTK
	Dr. Halápi Magdolna, egyetemi adjunktus, ELTE BTK
	Dr. Kimmel Magdolna, egyetemi adjunktus, ELTE BTK

2020

Acknowledgements

First, I would like to express my gratitude for those who have helped me complete this work. I would like to thank my supervisor, *Dr. Gergely Dávid*, for encouraging me to choose and work with such an eye-opening topic in the field of language testing. I would also like to thank him for his support and practical suggestions during all my studies. My appreciation also goes to the director of studies, *Dr. Dorottya Holló*, for her constant help, encouragement, and support during the whole PhD programme, as well as for her Research Seminars and for her help in writing up my dissertation. I am also very grateful to the programme director, *Dr. Krisztina Károly*, for her help and support in the administrative tasks and practical advice in completing the studies. Additionally, I would also like to express my thanks to *Dr. Tibor Vígh* for the help he provided me in understanding some German research articles and for the help he provided in the German data collection procedures.

I would also like to express my outmost and deepest gratitude to my family: my mother, *Ágnes*; my father, *Róbert*; and my sister, *Regina* for all their love, support, appreciation and devotion in every possible meaning of the word. I would like to thank them for making me who I am and making it possible for me to follow my dreams by providing me all the support for doing so. My deepest gratitude also goes to my partner, *Ágota*, for her love, appreciation, faith, patience, motivation, and immense knowledge she provided me in my life and during this whole project and showing me the light in the darkest moments.

Finally, I would like to thank one of my first English language teachers ever, *Ágnes Farnadi*, for teaching me English and for guiding and inspiring me in the beginning of my English language studies.

"We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run." — Roy Amara

Abstract

Technological development has a great influence on foreign language education, and using audio-visual material in foreign language classes is becoming a more and more widespread practice among language teachers. As the aim of language testing is to assess a skill in an artificial situation which successfully emulates the intended real-life situation, the changes in the real-world context of language use cannot be ignored by foreign language examinations. The influence of the increasing consumption of audio-visual material might be the most important in the case of the listening comprehension skill, so to maximise the authenticity and criterion-related validity of listening comprehension tests, the supplementation of the construct with audio-visual tasks might be taken into consideration. As using audio-visual material for testing listening comprehension and the reliability of such tests is an under-researched area both in the international and the Hungarian context, the aim of the present dissertation is to analyse whether including audio-visual tasks into the listening comprehension component of language examinations is necessary and desirable. This aim is fulfilled by designing a paper-based and a computer-based set of tasks for four different language proficiency levels (i.e., A2, B1, B2, C1) in two languages (i.e., English and German). The data collection was carried out with 140 participants, and their results on the two test versions were compared. In addition, two questionnaires were designed (i.e., a paper-based test questionnaire and a computer-based test questionnaire) which were intended to record the participants' opinions about the usefulness and necessity of the audio-visual material used in the tests. The results seem to suggest that the computer-based sets of tasks which contain the audio-visual tasks do not measure listening comprehension less reliably than the paper-based sets of tasks, and that the participants found the audio-visual material non-disturbing and especially useful for the lower language proficiency levels.

Keywords: listening comprehension, audio-visual comprehension, language testing, computer-based language testing

Table of Contents

1 Introduction 1
2 Theoretical background 6
2.1 Theoretical models and frameworks in language testing 6
2.2 Models of testing language competence 9
2.3 The construct of listening comprehension 17
2.4 Testing listening comprehension 24
2.5 The concept of validity in language testing 31
2.6 Using audio-visual materials in education 33
2.7 Testing audio-visual comprehension 38
3 Research questions 42
4 Research methods 43
4.1 Data collection and data analysis procedures 44
4.1.1 First phase: Task development 44
4.1.2 Second phase: Questionnaire development 60
4.1.3 Third phase: Conducting the pre-tests 65
4.2 Ethical considerations 72
5 Results and discussion 75
5.1 Research question 1: Do the paper-based sets of tasks and the computer-based sets of tasks measure listening comprehension in an equally reliable way? 75
5.1.1 Test results 76
5.1.2 Questionnaire results 103
5.1.3 Conclusion 107
5.2 Research question 2: Does the performance of the test-takers on the audio-visual-to-audio-only tasks differ from their performance on the audio-visual tasks? 107
5.3 Research question 3: Do the participants perceive the inclusion of audio-visual tasks as useful? 118
6 Conclusion 125
7 Limitations of the study and implications for further research 131
8 Pedagogical implications 134
9 Feasibility issues 138
References 141
Appendices 155

List of Tables

Table 1 Biographical Data of the Participants Solving the English Language Tasks in the First Phase 49
Table 2 Biographical Data of the Participants Solving the German Language Tasks in the First Phase 50
Table 3 Overall Listening Comprehension Scale 53
Table 4 Guidelines for Task Development 55
Table 5 Task Types Used in the Research Project....56
Table 6 Watching TV and Film Scale 57
Table 7 The Biographical Data of the Participants in the Second 63
Table 8 The Number of Participants Solving the English Tasks in the Third Phase....67
Table 9 The Number of Participants Solving the German Tasks in the Third Phase 68
Table 10 The Number of Tasks in the English Tests....68
Table 11 The Number of Tasks in the German Tests 68
Table 12 The Cronbach's Alpha Values of the Paper-Based Test Questionnaire Constructs 71
Table 13 The Cronbach's Alpha Values of the Computer-Based Test Questionnaire Constructs 71
Table 14 Cronbach's Alpha Values and Internal Consistency 77
Table 15 Reliability Measures of the English Paper-based Tests....77
Table 16 Item Facility Range....78
Table 17 A2 English Paper-Based Test: Item Facility Values and Point-Biserial Correlations 80
Table 18 B1 English Paper-Based Test: Item Facility Values and Point-Biserial Correlations 82
Table 19 B2 English Paper-Based Test: Item Facility Values and Point-Biserial Correlations 83
Table 20 C1 English Paper-Based Test: Item Facility Values and Point-Biserial Correlations 85
Table 21 Reliability Measures of the English Computer-Based Tests 86
Table 22 A2 English Computer-Based Test: Item Facility Values and Point-Biserial Correlations 87
Table 23 B1 English Computer-Based Test: Item Facility Values and Point-Biserial Correlations 88
Table 24 B2 English Computer-Based Test: Item Facility Values and Point-Biserial Correlations 89
Table 25 C1 English Computer-Based Test: Item Facility Values and Point-Biserial Correlations 91
Table 26 Reliability Measures of the German Paper-Based Tests 92
Table 27 A2 German Paper-Based Test: Item Facility Values and Point-Biserial Correlations 93
Table 28 B1 German Paper-Based Test: Item Facility Values and Point-Biserial Correlations 94
Table 29 B2 German Paper-Based Test: Item Facility Values and Point-Biserial Correlations 95
Table 30 C1 German Paper-Based Test: Item Facility Values and Point-Biserial Correlations 96
Table 31 Reliability Measures of the German Computer-Based Tests....97

Table 32 A2 German Computer-Based Test: Item Facility Values and Point-Biserial Correlations 98
Table 33 B1 German Computer-Based Test: Item Facility Values and Point-Biserial Correlations 99
Table 34 B2 German Computer-Based Tests: Item Facility Values and Point-Biserial Correlations 100
Table 35 C1 German Computer-Based Test: Item Facility Values and Point-Biserial Correlations 102
Table 36 Test Questionnaires: Descriptive Statistics....105
Table 37 Comparison of the Participants' Results on the Last Tasks in the English A2 Paper-Based and Computer-Based Tests....109
Table 38 Comparison of the Participants' Results on the Last Tasks in the English B1 Paper-Based and Computer-Based Tests....110
Table 39 Comparison of the Participants' Results on the Last Tasks in the English B2 Paper-Based and Computer-Based Tests....111
Table 40 Comparison of the Participants' Results on the Last Tasks in the English C1 Paper-Based and Computer-Based Tests....112
Table 41 Comparison of the Participants' Results on the Last Tasks in the German A2 Paper-Based and Computer-Based Tests....113
Table 42 Comparison of the Participants' Results on the Last Tasks in the German B1 Paper-Based and Computer-Based Tests....114
Table 43 Comparison of the Participants' Results on the Last Tasks in the German B2 Paper-Based and Computer-Based Tests....115
Table 44 Comparison of the Participants' Results on the Last Tasks in the German C1 Paper-Based and Computer-Based Tests....116
Table 45 Questionnaire Results: Necessity of the Video....119
Table 46 One-Way Analysis of Variance of the Necessity of the Video Regarding the English Paper-Based Tests 120
Table 47 Duncan Post Hoc Test for the Necessity of the Video Regarding the English Paper-Based Tests 121
Table 48 One-Way Analysis of Variance of the Necessity of the Video Regarding the English Computer-Based Tests 121
Table 49 Duncan Post Hoc Test for the Necessity of the Video Regarding the English Computer-Based Tests....122
Table 50 One-Way Analysis of Variance of the Necessity of the Video Regarding the German Paper-Based Tests....122
Table 51 Duncan Post Hoc Test for the Necessity of the Video Regarding the German Paper-Based Tests 123
Table 52 One-Way Analysis of Variance of the Necessity of the Video Regarding the German Computer-Based Tests....123
Table 53 Duncan Post Hoc Test for the Necessity of the Video Regarding the German Computer-Based Tests....124

List of Figures

Figure 1 Models, Frameworks, and Test specifications....7
Figure 2 Components of Communicative Language Ability 13

List of Appendices

Appendix 1A – Table A: Main characteristics of the A2 English language paper-based task set 155
Appendix 2A – Table B: Main characteristics of the B1 English language paper-based task set 156
Appendix 3A – Table C: Main characteristics of the B2 English language paper-based task set 157
Appendix 4A – Table D: Main characteristics of the C1 English language paper-based task set 158
Appendix 5A – Table E: Main characteristics of the A2 English language computer-based task set 159
Appendix 6A – Table F: Main characteristics of the B1 English language computer-based task set 160
Appendix 7A – Table G: Main characteristics of the B2 English language computer-based task set 161
Appendix 8A – Table H: Main characteristics of the C1 English language computer-based task set 162
Appendix 9A – Table I: Main characteristics of the A2 German language paper-based task set 163
Appendix 10A – Table J: Main characteristics of the B1 German language paper-based task set 164
Appendix 11A – Table K: Main characteristics of the B2 German language paper-based task set 165
Appendix 12A – Table L: Main characteristics of the C1 German language paper-based task set 166
Appendix 13A – Table M: Main characteristics of the A2 German language computer-based task set 167
Appendix 14A – Table N: Main characteristics of the B1 German language computer-based task set 168
Appendix 15A – Table O: Main characteristics of the B2 German language computer-based task set 169
Appendix 16A – Table P: Main characteristics of the C1 German language computer-based task set 170
Appendix 1B – Think-aloud tasks – the original Hungarian version and the English translation 171
Appendix 2B – Semi-structured interview – the original Hungarian version 172
Appendix 3B – Semi-structured interview – the English translation 174
Appendix 1C – Questionnaire in Hungarian about the Paper-Based Tests 176
Appendix 2C – Questionnaire about the Paper-Based Tests – English translation 179
Appendix 1D – Questionnaire in Hungarian about the Computer-Based Tests 182
Appendix 2D – Questionnaire about the Computer-Based Tests – English translation 185
Appendix 1E – Consent form in Hungarian 188
Appendix 2E – Consent form – English translation 189

List of Definitions of Frequently Used Terms

advocacy and participatory worldview – it intends to initiate change in a certain practice, and it promotes an open discussion about the issue in question (Creswell, 2009).

audio-visual comprehension – "the user watches TV, video, or a film and uses multi-media, with or without subtitles and voiceovers" (Council of Europe, 2018, p. 54). In contrast with the listening only activities, in case of audio-visual comprehension, the listener has to comprehend both audio and visual input.

audio-visual-to-audio-only (ATAO) task – in the case of the paper-based sets of tasks, the recording of the last task was modified by simply removing the visual material from the originally audio-visual recording.

cognitive validity – the relationship between the test performance and the criterion performance (Glasser, 1991).

computer-adaptive testing – "The function of an adaptive test is to present test items to an examinee according to the correctness of his or her previous responses. If a student answers an item correctly, a more difficult item is presented; and conversely, if an item is answered incorrectly, an easier item is given. In short, the test "adapts" to the examinee's level of ability. The computer's role is to evaluate the student's response, select an appropriate succeeding item and display it on the screen. The computer also notifies the examinee of the end of the test and of his or her level of performance." (Larson, 1989, p. 278)

computer-based set of tasks – one of the data collection instruments used in the present study. Depending on the language proficiency level, the computer-based set of tasks was a set of 4 or 5 listening comprehension tasks, where the last task of the set was an audio-visual task, while the rest of the tasks in the set were audio-only tasks. The computer-based set of tasks was administered to the participants on a digital platform. **In the present dissertation the terms *computer-based set of tasks* and *computer-based tests (CBT)* are used synonymously for the sake of convenience.**

construct-irrelevant variance – the specificities of the task that candidates have to solve in the test are irrelevant from the point of view of the construct (Messick, 1995).

construct underrepresentation – what candidates have to do in real-life tasks are not represented well enough in the testing situation (Messick, 1995).

construct validity – the results of the test mirror what the test is meant to measure (Messick, 1989).

criterion-related validity – the correspondence between the performance in real-life situations and the performance in the testing situation (Cohen, Manion & Morrison, 2000).

digital literacy – "the ability to locate, organise, understand, evaluate, analyse, create and communicate information using digital technologies" (Kaltura Report, 2015, p. 5).

framework – "a selection of skills and abilities from a model that are relevant to a specific assessment context" (Fulcher & Davidson, 2007, p. 36).

generation X – the people born between 1961 to 1981 (Strauss & Howe, 1997).

generation Z – people born after the year 1995 (Strauss & Howe, 1997).

listening comprehension – a listener receiving and processing "spoken input produced by one or more speakers" (Council of Europe, 2001, p. 65). During this process, besides the decoding of the message on a phonological, syntactic and word level, the listener's knowledge of the world and knowledge of schematic structures are also activated (Council of Europe, 2001, 2018).

model – "over-arching and relatively abstract theoretical descriptions of what it means to be able to communicate in a second language" (Fulcher & Davidson, 2007, p. 36).

paper-based set of tasks – one of the data collection instruments used in the present study. Depending on the language proficiency level, the paper-based set of tasks was a set of 4 or 5 audio-only listening comprehension tasks, where the last task of the set was an ATAO task created by removing the visual input from an originally audio-visual task. The paper-based set of tasks was administered to the participants in a printed out, paper-and-pen format. **In the present dissertation the terms *paper-based set of tasks* and *paper-based tests (PBT)* are used synonymously for the sake of convenience.**

principle of beneficence – the participants should gain some benefits from taking part in the data collection (Kubanyiova, 2015).

principle of justice – the requirement of fair distribution of research benefits (Kubanyiova, 2015).

principle of non-maleficence – it has to be ensured that "the research does not harm the subjects in any way (Cohen, Manion & Morrison, 2000, p. 71).

reliability – "consistency and replicability over time, over instruments and over groups of respondents" (Cohen, Manion & Morrison, 2000, p. 117).

test specifications – "generative explanatory documents for the creation of test tasks" (Fulcher & Davidson, 2007, p. 52).

test method facets – the method factors affecting test performance (Bachman, 1990).

1 Introduction

Language teaching and testing represent a constantly evolving field, where teaching and testing methodologies and instruments have to meet the changing demands of stakeholders. Listening comprehension is an area of language testing that is affected massively by the changes and challenges of both people's learning and perception orientations as well as technical development. With people becoming more and more visually oriented (Woolfitt, 2015) and audio and video playing equipment being more accessible for testing purposes, it seems relevant to study how audio-visual input affects test takers' performance in listening comprehension as opposed to audio-only input. The present dissertation, therefore, compares the performance of foreign language test takers on "traditional" listening tasks to their performance on audio-visual comprehension tasks, and analyses whether using audio-visual comprehension tasks has an effect on the test performance of the participants. Furthermore, the study also investigates whether it is necessary and desirable to include audio-visual materials in a testing situation. The terms *necessary* and *desirable* should be separated, as in the present dissertation, the term *necessary* is used in reference to the extent to which the real-world context and the methodology supports the legitimacy of extending the listening comprehension part of foreign language tests with audio-visual tasks. In contrast, the term *desirable* refers to whether the stakeholders involved in foreign language testing find it feasible and appealing to include audio-visual tasks in the listening component of foreign language tests.

To justify the need for introducing and examining an innovative method in language testing, it is necessary to briefly summarise trends that have affected foreign language teaching and testing in the past decades. The structuralist-behaviourist approach of language teaching and testing became an old-fashioned method by the early 1980s due to the arrival of communicative language teaching (Morrow, 1979). The development of both the

productive and receptive language skills has an important role in communicative language teaching. Communicative language teaching also puts the language learner in the centre of the learning process by declaring them to be an autonomous learner who is responsible for their own learning progress (Bárdos, 2005). Structuralist types of activities (e.g., drill types of exercises) were replaced by interactive and problem-solving oriented activities and tasks. The role of the teacher is to initiate the context to these interactive tasks to make the learning context more communicative. Language teachers, by their own account, also try to design their language classes to be as communicative as possible (Bárdos, 2005). Therefore, by today, the concept of communicative language teaching is a widely shared teaching approach in foreign language education.

Because of the influence of communicative language teaching, the traditional structuralist-behaviourist approach in language testing became outdated as well. As a result, language tests had to be redesigned in a way to follow the principles of communicative language teaching (Morrow, 1979). In the past decades, language testing professionals, therefore, have attempted to redesign, with more and sometimes with less success, their language test tasks in a way to make the artificial language testing situation more communicative and reflective of the real-world context. In contrast with the structuralist-behaviourist approach, which focused on testing language competence instead of performance, communicative language testing aims to assess the performance of the test-taker in a foreign language through spoken and written language production (McNamara, 1996; Morrow, 1979).

Communicative language teaching and communicative language testing emerged in the 1970s (Morrow, 1979); therefore, they were created in a vastly different social context from today's environment. As a result of the rapid technological advancement experienced in the past 40 years, the instruments available to aid language teaching have substantially

changed. Findings of recent studies on the language learning habits of students both in-class and outside the classroom suggest that the use of technology, such as watching videos and films in the target language and using language learning applications designed for language learning purposes are very popular among language learners (Bates, 2015; Fransen, 2015; Greenberg & Zanetis, 2012; Woolfitt, 2015). Such technological inventions were not available at the dawn of communicative language teaching; however, research about teaching practice suggests that there is a strong attempt in foreign language teaching to adapt to the changing context (Greenberg & Zanetis, 2012; Guo, Kim, & Rubin, 2014; Woolfitt, 2015). Similar efforts can be observed in the field of language testing as major language examinations, like Cambridge and TOEFL, already offer the opportunity to take the examination in a computer-based format (Cambridge Assessment, 2019; ETS TOEFL, 2019). However, at the time of conducting the present research study, such practice was still not available in the case of most of the smaller language examinations, especially in the Hungarian context.

Keeping the context-embedded principle of communicative language teaching in mind, the tasks used in language tests have to be constantly updated and improved to match the changing real-world context. Computer-based language testing could especially aid the improvement of the testing of listening comprehension by adding new task types which would more authentically represent real-world listening activities. It might especially become problematic that the use of the audio-visual materials is not widely applied in language tests because it can result in the listening comprehension construct being underrepresented (Messick, 1995). As consuming audio-visual media in the form of TV programmes and online videos has become part of people's everyday life, those language tests which intend to adequately simulate circumstances and problems a language user might encounter in a real-life situation should probably include audio-visual materials.

Another reason for considering the revision of the task types used for language testing is the fact that the main approaches and ways of communicative language testing were laid down in the 1970s with a different generation from today's generation in mind. Taking the works of Strauss and Howe (1997) and Howe and Strauss (2007) into consideration, the beginnings of communicative language teaching and language testing can be placed to the time when the members of *Generation X* were going to school. *Generation X* refers to the people born between 1961 and 1981 (Strauss & Howe, 1997). Howe and Strauss (2007) describe the social environment of *Generation X* as crucially different from that of today's generation labelled *Generation Z*. The term *Generation Z* applies to people born after the year 1995, and they form the generational cohort which is considered to have had ready access to technological advancements such as smartphones, computers and the Internet from their early childhood (Howe & Strauss, 2007). In contrast, members of *Generation X* did not have access to such features during their childhood. Even though the age range belonging to the term *Generation Z* and the generation theory (Strauss & Howe, 1997) itself are widely disputed concepts among researchers (Combi, 2015; McCrindle & Wolfinger, 2014; Palfrey & Gasser, 2008), and this theory was designed with the American social context in mind — so the Hungarian social context applicable for the different generations could show major differences — it cannot be debated that the technological tools available for language learning and the use of technology in general have considerably changed in the past 40 years.

Despite the fact that the amount of available audio and audio-visual materials is larger than ever and attempts have been made to incorporate them into language education (Bates, 2015; De Vera & McDonnell, 1985; Greenberg & Zanetis, 2012), the depth of research data on using audio-visual materials in language teaching and language testing is still insufficient. Research on the assessment of audio-visual text comprehension has already

been carried out (Kellerman, 1990; Ockey, 2007; Raffler-Engel, 1980; Sueyoshi & Hardison, 2005). However, the amount of research is sparse, in fact non-existent in the Hungarian context, and the research results are incongruent with each other. Therefore, further research is needed on this issue. To contribute to this research niche, the present dissertation piloted 16 listening comprehension tests designed for four different language proficiency levels, namely, A2, B1, B2, C1 (Council of Europe, 2001, 2018) administered in two different formats (i.e., paper-based test and computer-based test) and in two different languages (i.e., English and German). Furthermore, the research study also investigated (1) the reliability of the developed tests, (2) the performance of the students on the different test formats, (3) the students' perceptions of the different test features, and (4) the necessity and desirability of including audio-visual materials in a testing situation.

For the sake of a logical presentation of the research study, this dissertation is structured as follows: First, a review of the relevant literature is provided in Chapter 2 (p. 6) to provide the theoretical background concerning the topics of language testing, testing listening comprehension, validity in language testing, audio-visual comprehension, and whether including an audio-visual component could enhance measuring listening comprehension. Chapter 3 (p. 42) presents the research questions, then Chapter 4 (p. 43) discusses the research methods used in the present study, providing details about the research instruments, the participants, the data collection, and the methods of data analysis. The results and the discussion of the data analyses are presented in Chapter 5 (p. 75). Finally, conclusions are drawn in Chapter 6 (p. 125), and Chapter 7 (p. 131), 8 (p. 134), and 9 (p. 138) discuss the limitations of the study, the pedagogical implications, and the possible feasibility issues.

2 Theoretical background

The aim of the following section is to provide an extensive overview of the research conducted in the topic of testing listening and audio-visual text comprehension. In order to do so, the following section is divided into seven sub-sections: first, the theoretical models and frameworks underpinning language testing are presented (p. 6); in the second section, the communicative language competence models are discussed (p. 9); thirdly, the construct of listening comprehension is analysed (p. 17); the fourth section, presents how listening comprehension is tested (p. 24); the fifth section introduces the concept of validity in language testing (p. 31); finally, the sixth and seventh sections discuss the construct of audio-visual comprehension (p. 33) and how it could be tested (p. 38).

2.1 Theoretical models and frameworks in language testing

Language testing plays an important role in nowadays' education in Hungary as having a B2 level language certificate in a foreign language is a pre-requisite for college students to receive their BA or BSc degrees (OM Rendelet [Education Decree], 2006). In the light of this, valid and reliable language testing is crucial for those students who would like to participate in the Hungarian tertiary education system. Therefore, it has to be made sure that each language skill and competence is measured by the language examinations as accurately as possible.

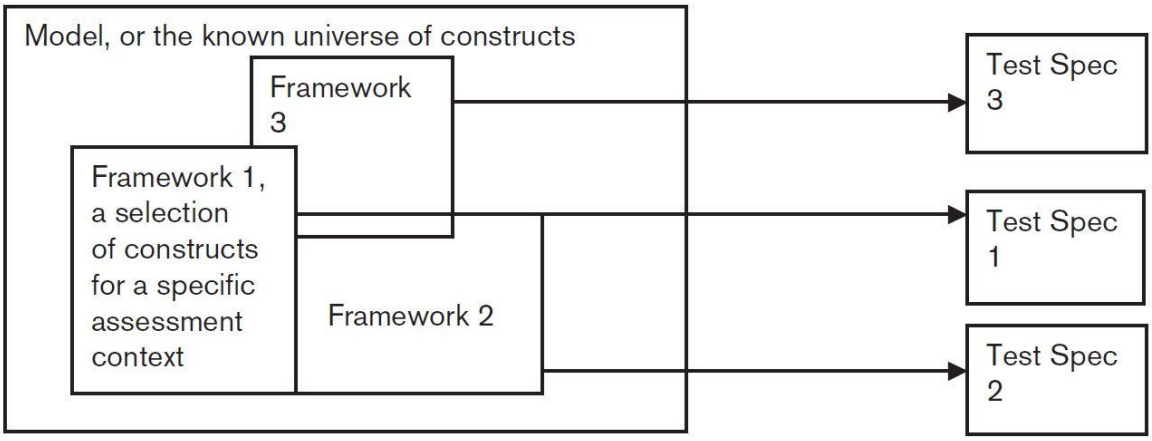
Based on the findings of previous research (Gósy, 2000; Kuang-yun, 2007; Petőné Honvári, 2014; Szabó & Nikolov, 2013), listening comprehension appears to be a rather problematic construct in language teaching and examinations. Therefore, the aim of this dissertation is to examine whether it is desirable and advisable to expand the listening comprehension component of language tests with tasks targeting audio-visual text comprehension. To be able to do so, first, the most important concepts of language testing

have to be discussed so the present section provides the definition of basic language testing terms and it discusses the theoretical background of assessing different language skills.

In test development, three different layers of terms have to be distinguished: *models*, *frameworks* and *test specifications*. First, the concepts of models and frameworks have to be distinguished. Regarding the distinction between models and frameworks, the definitions are usually vague. The definitions can be especially confusing because the two terms are sometimes used interchangeably in the literature. To overcome this issue, the present dissertation follows the definitions designed by Fulcher and Davidson (2007) based on the work of Chalhoub-Deville (1997) (see Figure 1). According to the definitions of Fulcher and Davidson (2007), *models* are "over-arching and relatively abstract theoretical descriptions of what it means to be able to communicate in a second language" (p. 36), whereas *frameworks* can be defined as "a selection of skills and abilities from a model that are relevant to a specific assessment context" (p. 36). *Test specifications*, however, are "generative explanatory documents for the creation of test tasks" (Fulcher & Davidson, 2007, p. 52).

Figure 1

Models, Frameworks, and Test Specifications (Fulcher & Davidson, 2007, p. 37)



Therefore, it can be concluded that a model is a relatively abstract description of language knowledge and use; whereas test specifications are concrete guidelines about the structure, requirements and design of a particular test. The connection between these two is created by the framework, which describes those aspects of a model which are relevant to certain language use domains (Fulcher & Davidson 2007). This shows that models serve as the core of language testing. However, models handle competence and performance in an extensively broad way and they do not account for specific contexts. For this reason, frameworks have to be specified by taking the audience of the test, the use of scores, and the performance conditions of the test takers (Fulcher & Davidson, 2007).

The test specifications are written on the basis of the framework. According to Fulcher and Davidson (2007), test specifications are explanatory documents which provide a detailed description of the tasks in a test. Their role is to ensure test equivalence, which means that new tasks in a test have the same level of difficulty and testing objective as previous ones. For this reason, test specifications have two main elements: samples of tasks and guiding language (Fulcher & Davidson, 2007). Therefore, the role of frameworks and that of test specifications are to make language competence context specific and accessible for testing.

The test specifications also contain a detailed description of the constructs measured by the test. According to McNamara (2000) *test constructs* refer "to those aspects of knowledge or skill possessed by the candidate which are being measured" (p. 13). To be able to appropriately define the constructs, first the test's definition of knowledge and its performance criteria have to be established. These influence every aspect of the language test, from the structure of the test to the interpretation of the test scores (McNamara, 2000).

Models of communicative competence and performance serve as the basis for large-scale language testing at present. However, language competence has been interpreted

in several different ways throughout the last century. Therefore, the following section provides an overview of the most influential models in language testing, paying special attention to the development of the models of communicative competence.

2.2 Models of testing language competence

The first theories about language competence were based on the notions of structuralist linguistics, and they viewed language knowledge as a set of systems. The most influential advocate of this view was Lado (1961), who promoted discrete point testing for testing language knowledge. Discrete point testing focused on testing the examinee's grammar, vocabulary and pronunciation knowledge in an isolated and decontextualized way. The testing of these skills was carried out mostly with the help of isolated sentences and multiple-choice questions (Lado, 1961). According to McNamara (2000), attempts of integrated testing of the performance were also made in the 1960s, so discrete point tests were also supplemented with the testing of the four macro-skills; however, listening, speaking, reading and writing skills were also tested in isolation. This trend is labelled as the psychometric-structuralist period (McNamara, 2000).

As the discrete point testing focused only on the knowledge of the linguistic system without a context and it failed to assess language knowledge used for communication, in the 1970s, a need for a more communication-oriented way of language testing emerged. This resulted in the first considerations of using integrative testing (McNamara, 2000). However, creating integrative tests proved to be more expensive, more difficult to score and could lead to potential unreliability because those who scored the tests could easily disagree about the acceptable answers. Oller (1976) developed an interpretation of language knowledge, which seemed to offer a different alternative, and which later came to be known as the *Unitary Competence Hypothesis* (Oller, 1976). According to the *Unitary Competence Hypothesis* (Oller, 1976), language competence has two main components: *real time language*

comprehension and *pragmatic mapping*. The first component refers to understanding language in communication situations involving listening and speaking, whereas *pragmatic mapping* refers to the use of one's formal systemic knowledge about the language to understand contextualised meaning (Oller, 1976). The *Unitary Competence Hypothesis* (Oller, 1976) claims that the test performance of a test taker depends on being able to combine grammar, vocabulary, contextual and pragmatic knowledge during the test. Therefore, gap-filling tests such as cloze tests were considered to be perfectly appropriate for testing the necessary skills and to substitute for the more expensive listening, speaking, reading and writing tests. Their most compelling features were their lack of difficulties in construction and scoring (McNamara, 2000). Even though the *Unitary Competence Hypothesis* had some merits in describing language performance, it was later proved that Oller (1976) used inappropriate methods of data analysis in his study, and that cloze tests are not appropriate for testing communicative skills (McNamara, 2000).

In the 1970s, another trend in the interpretation of language knowledge also seemed to be emerging. Chomsky, who considered language competence as a native speaker's knowledge of the language, provided one of the first notable discussions of language competence. In contrast with Oller (1976), who was concerned with the unity of the language competence, Chomsky (1965) was interested in the connection between language competence and performance. He named the concept of language competence *Universal Grammar* (Chomsky, 1965). This idea was further developed by Hymes (1971, 1972), who divided linguistic competence into four different components, namely *knowledge of possibility, feasibility, appropriateness* and *attestedness*. The component called *knowledge of possibility* is considered to be roughly the equivalent of Chomsky's *Universal Grammar*, and it contains everything the speaker knows about the grammar rules of the language. *Feasibility* refers to the information load the brain is able to comprehend and process. For

instance, the difficulty of processing multiple recursive forms is related to the *feasibility* component of language competence (Hymes, 1971, 1972). The other two components, *appropriateness* and *attestedness* are not present in Chomsky's model (Chomsky, 1965). According to Hymes (1971, 1972), *appropriateness* refers to the ability to meet the contextual and social requirements of language use in a situation (e.g., being able to decide whether formal or informal language use is more appropriate in a particular situation), whereas *attestedness* is the correct knowledge of idiomatic expressions in a language (e.g., the correct idiom in English is "ups and downs" and never "downs and ups"). These two components form the language users' sociolinguistic competence, and their inclusion into Hymes' (1971, 1972) language competence model marks the beginning of the era dominated by the communicative competence theory (McNamara, 1996).

Concentrating on the communicative focus of language competence, several models were proposed to describe the elements of language use. The first and most notable communicative competence model was created by Canale and Swain (1980). Their main aim with defining communicative competence was to support second language (L2) teaching by providing a model based on which a valid and more reliable measurement of the language skills could be developed (Canale & Swain, 1980). Their model divides language competence into two main categories: *communicative competence* and *actual communication*. Communicative competence contains grammatical competence (i.e., the knowledge of grammatical, lexical, morphological, syntactic, semantic and phonological rules of a language), sociolinguistic knowledge (i.e., being aware of the sociocultural rules connected to discourse and language use), and strategic competence (i.e., the ability to overcome communication problems and difficulties). In contrast, actual communication refers to demonstrating one's language knowledge through performance (Canale & Swain, 1980). The grammatical competence component of Canale and Swain's (1980) model is

actually the same as Chomsky's notion of linguistic competence, and sociolinguistic knowledge and strategic competence were created by dividing the sociolinguistic competence component of Hymes' (1971, 1972) model. This model makes a clear distinction between communicative competence and communicative performance, and Canale and Swain (1980) argue that assessment of the language knowledge has to be done with tests which access communicative competence through tasks resembling language use in real life situations. Therefore, this model is highly relevant for the field of language testing.

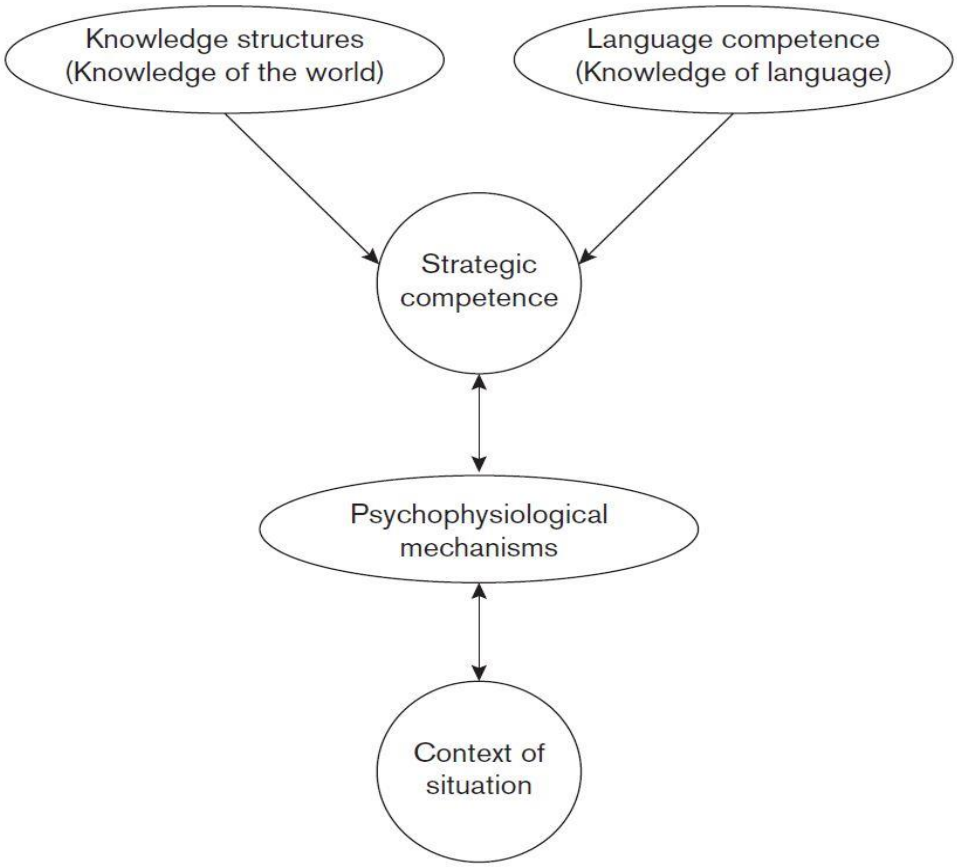
The communicative competence model of Canale and Swain (1980) was expanded by Canale himself (1983a, 1983b) as he added the concept of *discourse competence* to the component of sociolinguistic knowledge, and he re-interpreted the actual communication component as "the realization of such knowledge and skill under limiting psychological and environmental conditions such as memory and perceptual constraints, fatigue, nervousness, distractions and interfering background noises" (Canale, 1983a, p. 5). This novel view of the components resulted in a new definition of communicative competence: "communicative competence refers to both knowledge and skill in using this knowledge when interacting in actual communication" (Canale, 1983a, p. 5). As a result of the new definition, Canale (1983a, 1983b) handled communicative competence separately from actual communication because performance in a concrete situation (i.e., actual communication) was considered to be the manifestation of the underlying knowledge and skills (i.e., communicative competence). The other components of the models were also re-interpreted: sociolinguistic competence was restricted to the knowledge of sociocultural roles, whereas the rules of discourse were contained in the discourse competence.

This revised version (Canale, 1983a, 1983b) of the Canale and Swain communicative competence model (1980) served as the basis of all further language competence models.

The next prominent model is Bachman's (1990) model of *communicative language ability* (CLA) (see Figure 2). Building on previous language competence models Bachman's (1990) CLA represents a more detailed description of communicative competence. The components of CLA (Bachman, 1990) are language competence, strategic competence and psychophysiological mechanisms. Language competence contains the knowledge of the language, strategic competence means the ability to apply the components of language competence in a certain context, and psychophysiological mechanisms make the physical execution of language possible. Bachman (1990) claimed that strategic competence is also influenced by the language user's knowledge of the world. Compared to the previous models, Bachman's (1990) CLA also creates a clear differentiation between the notions of knowledge and skills.

Figure 2

Components of Communicative Language Ability (Bachman, 1990, p. 85)



CLA (Bachman, 1990) was later revised and restructured by Bachman and Palmer (1996). They introduced the affective (i.e., non-cognitive) schemata, re-defined the strategic competence component as metacognitive strategies, and re-named the language user's knowledge of the world as topical knowledge. Bachman and Palmer (1996) defined affective schemata as "affective or emotional correlates of topical knowledge" (p. 65), namely, "the memories or past experiences that determine whether an individual will engage with a particular task" (Fulcher & Davidson, 2007, p. 45).

Similarly to Bachman's (1990) and Bachman and Palmer's (1996) models, other ones such as Celce-Murcia, Dörnyei and Thurrell's (1995) model were also based on the reinterpretation of Canale's (1983a, 1983b) model of communicative competence. Celce-Murcia, Dörnyei and Thurrell (1995) initiated the addition of *actional competence* to the components proposed by Canale (1983a, 1983b). However, Celce-Murcia, Dörnyei and Thurrell's (1995) model is not going to be considered in the present dissertation as it is primarily designed for a teaching context and classroom setting, and it is unfit for language testing purposes. In the field of language testing Bachman's (1990) model and its revised version (Bachman & Palmer, 1996, 2010) are still the most comprehensive and most detailed model because they combine the information for the constructs from several different applied linguistics fields.

One of the newest interpretations of the language competence is provided by the *Common European Framework of Reference* (CEFR) (Council of Europe, 2001). The CEFR (Council of Europe, 2001) proposes an action-oriented approach towards language knowledge, and it views language use as an act carried out by a language user in order to accomplish tasks in certain social contexts. Therefore, it defines language use as follows:

> Language use, embracing language learning, comprises the actions performed by persons who as individuals and as social agents develop a range of competences,

> both general and in particular communicative language competences. They draw on the competences at their disposal in various contexts under various conditions and under various constraints to engage in language activities involving language processes to produce and/or receive texts in relation to themes in specific domains, activating those strategies which seem most appropriate for carrying out the tasks to be accomplished. The monitoring of these actions by the participants leads to the reinforcement or modification of their competences. (Council of Europe, 2001, p. 9)

Based on this definition, the main components of the language competence according to the CEFR (Council of Europe, 2001) are the language user's *general competences*, *communicative language competences*, the *context*, the *language activities*, the *language processes*, the *language use domains*, the *text*, the *task solving strategies*, and the *tasks* themselves. The *general competences* involve the language user's declarative knowledge resulting from academic and empirical knowledge; skills or know-how about executing procedures; existential competence, which contains an individual's attitudes, personality traits and characteristics; and the individual's ability to learn. *Communicative language competences* have three components: linguistic elements, sociolinguistic elements and pragmatic elements. The influence of previous communicative competence models is clearly visible because the linguistic component is concerned with the language user's knowledge about the formal systemic features of the language, the sociolinguistic component describes the knowledge about the sociocultural aspects of a language, and the pragmatic component contains the knowledge of speech acts and scenarios of interactions. *Language activities* refer to the interactions which activate the language user's communicative competences, the *context* is the collection of situational features which the communication is ingrained into, whereas *domains* refer to the different sectors of language use, namely, public, personal,

educational and occupational domain. Regarding *tasks*, *strategies* and *texts*, *tasks* are purposefully carried out actions to solve a problem or achieve a goal, and as these actions are not automatic, the language user has to use certain *strategies* to achieve these results (i.e., intentional and regulated actions). During these processes, the language user has to comprehend and produce oral or written *texts*. The aforementioned components of language competence are considered to be intertwined in every instance of language use (Council of Europe, 2001).

As the CEFR (Council of Europe, 2001) and its later revised version (i.e., Council of Europe, 2018) serve as the basis for language test design in Europe, the present study also considers the CEFR (Council of Europe, 2001) as its theoretical basis with regards to language competence. As it is discussed in further detail in the methods section (see section *4.1.1.2,* p. 51), the listening and audio-visual comprehension tasks used in the present study were also calibrated based on the theoretical background and requirements described by the CEFR (Council of Europe, 2001). Despite the fact that the scales of CEFR (Council of Europe, 2001) were revised in 2016-2017, this section used the 2001 version of the document as a reference for the theoretical background. This choice was made because the tasks used for data collection were created before the new CEFR descriptors (Council of Europe, 2018) were published. Furthermore, the theoretical background presented in CEFR (Council of Europe, 2001) was not modified in the revised edition so using the new edition would have had no effect on the task design.

The overview of the language competence models suggests that the approach to the components of language competence and to the ways of testing have gone through major changes in the past decades. The following section focuses on the discussion of the listening comprehension construct in detail.

2.3 The construct of listening comprehension

One of the constructs traditionally measured by language tests is listening comprehension. Listening comprehension plays an important role in humans' life. It is part of people's everyday face-to-face, telephone and online conversations or when they watch or listen to pre-recorded materials on TV, radio or the Internet. Based on estimations, people spend at least 50% of communication listening (Wagner, 2014). In fact, the understanding of speech is of primary importance not only in verbal communication but also in language education, as good listening comprehension both provides input for the learner and opens the way to direct face-to-face communication in a foreign language.

The term listening comprehension has been defined in several different ways. One of the basic and most concise definitions of the term is provided by Rost (1990):

> Understanding spoken language is essentially an inferential process based on a perception of cues rather than a straightforward matching of sound to meaning. The listener must find relevant links between what is heard (and seen) and those aspects of context that might motivate the speaker to make a particular utterance at a particular time. (p. 33)

Based on this definition, speech comprehension requires the listener to decode utterances, which is why it is necessary to discuss how speech perception can happen. According to Marslen-Wilson and Tyler (1980), the acoustic characteristics of sounds (e.g., length and loudness) help the listener to decode the different speech signals from the stream of sounds. Besides the acoustic characteristics, speech perception also depends on time because understanding speech signals requires some time to be processed (Brazil, 1983; Chafe, 1980, 1982; Kreckel 1981). That is the reason why it might be difficult to follow and understand someone who is jabbering.

Decoding the utterances also requires identifying phonemic units, namely, phonemes, which are considered to be the smallest units of speech (Chomsky & Halle, 1968). The realisation of phonemic units is influenced by the co-articulation of sounds within a word and by mapping the abstract phoneme to one of its variables, for example, the phoneme /l/ is pronounced differently in the words *file* and *life*.

However, in real-time listening comprehension the listener does not only have to identify the physical characteristics of sounds and derive the abstract phonemes into their variations, but they also have to use their pragmatic knowledge to understand the meaning of the words, and to keep all this meaning in their short-term memory at the same time (Berg, 1987; Bregman, 1978; Buck, 2001). In fact, from the communicative language teaching and testing point of view, understanding the meaning is more important than the psycholinguistic processes applied during listening. In the communicative language teaching and testing approach, there are different language competence models (Canale, 1983a, 1983b; Canale & Swain, 1980; Celce-Murcia, Dörnyei & Thurell, 1995; Council of Europe, 2001; Hymes, 1971, 1972) with the help of which researchers have attempted to map the elements of language competence. These models concentrate more on the various social and pragmatic elements, or in other words, how language is used in context, rather than how the language is processed in the human mind. Therefore, the complex nature of the listening process emerges both from the psycholinguistic processes and the verbal communication (i.e., social and pragmatic contexts).

It is not surprising that investigating listening comprehension abilities and the nature of listening comprehension have always been within the scope of theoretical and applied linguistics research (Buck 2001; Buck & Tatsuoka, 1998; Lund, 1991; Richards, 1983; Valette, 1977; Weir, 1993). One of the earliest models describing the process of listening comprehension ability is based on empirical evidence provided by Clark and Clark (1977).

The model includes four stages of the psychological procedures underlying verbal communication. The stages of the procedures are as follows: first, the listener perceives the speech and attaches it to phonological representations stored in the working memory; the second step is the identification of the content and function of the phonological representations organised into constituents; based on the identified constituents the underlying propositions are organised into a hierarchical representation; as the last step, the listener stores the identified constituents and eventually, after some time, by forgetting the exact wording of the constituent, he only remembers the meaning. The main criticism this model received is that it disregards the context in which the speech is produced and it presupposes that the understanding of spoken language can only happen in this order (Rost, 1990).

A more elaborate and more theoretical listening comprehension model was developed by Demyankov (1983) who accounted for almost all aspects of speech comprehension even including acquiring a linguistic framework of the language, hypothesis testing of what is being heard, the illocution of the utterance (i.e., the speaker's intention), and the tone of the message. However, because of its highly theoretical nature, the model failed to realise real-time speech comprehension and the way in which ordinary conversations happen between interlocutors.

Another notable example is Richards' (1983) taxonomy of listening comprehension skills, which divided the process of listening comprehension into different micro-skills related to conversational listening and academic listening. Micro-skills related to conversational listening include such abilities among others as recognizing stress patterns, distinguishing word boundaries, and detecting sentence constituents. The academic listening micro-skills include identifying purpose and scope of a lecture, inferring relationships, and recognizing markers of cohesion (Richards, 1983). The main criticism against Richards's

(1983) taxonomy was that it does not provide a clear definition of how the micro-skills create the process, or how these components can be organised into a systematic hierarchy (Dunkel, Henning & Chaudron, 1993).

According to Buck (2001), listening comprehension is a complex skill, which necessitates the listener to be able to extract information and interpret it in context. Therefore, arriving at the correct interpretation of the information requires not only understanding linguistic features but also the correct interpretation of the context. Buck (2001) claims that there are three types of knowledge contributing to listening comprehension: language knowledge, world knowledge, and the ability to create mental representations of meaning. Therefore, listening comprehension is influenced by language competence (i.e., grammatical, sociolinguistic, pragmatic, and discourse knowledge) and the so-called strategic competence, which involves the use of cognitive and metacognitive strategies (Buck, 2001).

Even though there are several different models and taxonomies of listening comprehension with different components in them, most listening comprehension models agree that the listening comprehension process can be divided into two stages (Buck, 1991; Conrad, 1989; Lund, 1991; Rost, 1990; Weir 1993). The names of these two stages differ in the different research studies; however, they all seem to agree that, regardless of the label used for naming it, listening comprehension involves a first stage of lower order processes and a second one of higher order processes. These processes can generally be labelled *bottom-up* and *top-down* processing respectively. According to Brindley (1998), the first stage of listening comprehension involves the understanding of the information of the input literally (i.e., bottom-up processing), while the second stage involves forming critical evaluations about this information (i.e., top-down processing). Kelly (1991) defines the two stages similarly by stating that the first step is receiving the sound input and starting to

process it. During the second step, the sound input is given a meaning. The bottom-up and top-down processing are defined to be circular processes with the two stages happening simultaneously, and comprehension is achieved when the two stages provide enough information for the listener (Kelly, 1991).

One of the newest listening comprehension models created by Field (2009) also takes a similar approach to the process. This is one of the most detailed listening comprehension models, and according to Field (2009), the process of listening comprehension can be divided into two different sets of processes: decoding processes and meaning-building processes. During the decoding processes, the listener interprets the speech signals on the level of phonemes and syllables first, then a word level interpretation of the input follows through lexical segmentation and the activation of word networks in the listener's mind. The word-level processing is followed by syntactic parsing, where the syntactic structures are processed, and inferences are drawn based on them. Finally, the intonation, stress, pitch, loudness, speech rate, and accents are processed. In comparison, during the meaning-building processes, the listener interprets the meaning of the input and expands the information already received during the communication with that said input, by first interpreting the possible meanings of the words in context. After that, the context appropriate meaning is attached to the syntactic structures used by the speaker, and the appropriate inferences are drawn from them. The contextually appropriate meaning is also attached to the intonation, and the contextual and schematic knowledge are applied to the interpretation. During the final steps of meaning creation, inferencing is used to unfold implicit meanings, reference connections are recognised, the relevance of the input is considered, possible incoherences are handled, and the new pieces of information are integrated with the previously communicated ones. As the last step of the process, the discourse representation is created, revised or updated (Field, 2009).

Although Field (2009) described similar processes to the previously discussed models, he refrained from using the terms *top-down* and *bottom-up processing* in the traditional sense. He claimed that in the interpretation of his model, *bottom-up processing* refers to "building small units into larger" and *top-down processing* means "the influence of larger units when identifying smaller ones" (Field, 2009, p. 132). Field (2009) suggested that these two processes do not always necessarily occur in a specific order, and they can serve multiple different purposes, such as filling in information gaps in understanding or supplementing on decoded information.

The two stages of comprehension are not always used to the same extent either. When the words of the input are predictable, bottom-up processing is used to a lesser extent. Therefore, research suggests that beginner foreign language learners might have to rely on bottom-up processing more than their higher proficiency peers (Kelly, 1991). Evidence for this idea has been found by several researchers (Brown, 1986; Buck, 1994; Conrad, 1985; Hansen & Jensen, 1994; Shohamy & Inbar, 1991; Wu, 1998). For example, Hansen and Jensen (1994) compared the listening test results of candidates with different levels of language proficiency, and they found that learners who had lower language proficiency levels struggled with answering global questions more than their higher proficiency peers. Answering broader questions requires top-down processing rather than relying verbatim on the input. In comparison, when the same candidates had to answer questions relying on bottom-up processing by finding the verbatim answers in the input, they had considerably less difficulty answering the questions (Hansen & Jensen, 1994).

Even though none of the aforementioned comprehension models are empirically validated, the fact that researchers arrived at the same conclusions about the stages of listening comprehension independently from each other, and the fact that candidates with different language proficiency levels seem to struggle with different types of comprehension

problems, make the two-stage view of listening comprehension highly credible (Wagner, 2002).

As the tasks used in the present study were calibrated to match the requirements of the CEFR (Council of Europe, 2001), the discussion on the views about the construct of listening comprehension should be finished with the examination of the CEFR's approach. According to the CEFR (Council of Europe, 2001), listening comprehension is defined as a listener receiving and processing "spoken input produced by one or more speakers" (p. 65). During this process, besides the decoding of the message on a phonological, syntactic and word level, the listener's knowledge of the world and knowledge of schematic structures are also activated (Council of Europe, 2001).

Based on the CEFR (Council of Europe, 2001), listening comprehension is activated by different language activities, and it involves one or a combination of four processes, namely, *reception*, *production*, *interaction*, and *mediation*. Reception and production are considered to be primary processes, and reception can even occur without the presence of two individuals participating in the communication, for instance, when consuming media. In contrast, interaction can only occur with the participation of at least two individuals. In case of the interaction, the participants usually alternate between reception and production; however, CEFR (Council of Europe, 2001) also considers that fact that there is probably an overlap between the two processes because while listening to the speaking partner, the listener might also already start thinking about his or her answer. The act of mediation refers to situations where direct communication is impossible for the speaking partners. Therefore, mediation involves such processes as translation, interpretation, summarisation and paraphrasing (Council of Europe, 2001).

According to the CEFR (Council of Europe, 2001), typical listening situations involve listening to public announcements, listening to media recordings, listening as a

member of a live audience, or listening to overheard conversations. In any of these situations, the listener might be looking for different types of information, such as the gist of the input, the main ideas of the input, or specific details. The listener can also focus on gaining a detailed understanding, finding implications, or understanding the speaker's attitude towards the listener and the topic (Council of Europe, 2001).

The CEFR scales (Council of Europe, 2001) were revised in 2018 (Council of Europe, 2018), but from the point of view of the present study, no relevant modifications of the listening comprehension scales were proposed. Furthermore, the data collection for the present study started at the beginning of 2017 so at that time only the 2001 version of the CEFR (Council of Europe, 2001) was available for task design purposes. These are the reasons why the present theoretical overview decided to focus on the 2001 version of the CEFR (Council of Europe, 2001) instead of the 2018 version (Council of Europe, 2018).

2.4 Testing listening comprehension

The nature of listening comprehension is not only complex because of the different components of the listening construct, but also because of test method facets (Bachman, 1990), or in other words, the factors affecting test performance. The effects of these factors cannot be disregarded because test scores serve as evidence of the test taker's language competence (Fulcher & Davidson, 2007). A test score, however, might not only include information about the test taker's language competence but also about other factors, such as the temperature of the room the test is taken in, the behaviour of the invigilators, or the quality of the recordings – in case of a listening test – which can all affect and sometimes distort the evidence of the test taker's language competence.

The effects of test method facets have been researched extensively (Carroll, 1968; Clark, 1972; Cohen, 1980; Morrow, 1977; Weir, 1983), but Bachman's (1990) explanations provide a deeper understanding of pervious frameworks; thus, the current study uses his

approach of looking at test method facets. According to this categorisation, there are five facets which can affect test performance. They are as follows:

1) the testing environment,
2) the test rubric,
3) the nature of the input the test taker receives,
4) the nature of the expected response to the input, and
5) the relationship between the input and the response (Bachman, 1990).

As far as the *testing environment* is concerned, test takers' performance can be highly affected by those physical and environmental conditions in which the test is taking place. These facets include, for example, the test takers' familiarity with the place where they sit the test, the equipment (e.g., paper-based test and computer-based test) they need to use for completing the test, the people (e.g., invigilators, examiners, and fellow examinees) they need to interact with on the day of the test, and the time (e.g., morning and afternoon tests) and physical conditions (e.g., temperature and lightning) of the test (Bachman, 1990). Regarding the present study, one of the most important of these components is the use of equipment because the participants completed both a paper-based and a computer-based test. Therefore, the effect of such means of test completion might affect the evidence more than it is expected.

In general, a computer-based language test can be defined as "an integrated procedure in which language performance is elicited and assessed with the help of a computer" (Noijons, 1994, p. 38). According to Noijons (1994), the integration of the computer platform can be done at three different phases of the language examination: in test generation, in interacting with the candidates, and in evaluating the candidates' responses. At each one of these phases, using a computer can have different advantages and disadvantages (Chapelle & Douglas, 2006).

Taking the first phase into consideration, according to Sulaiman and Kahn (2019), the most typical use of a computer in generating the tests is connected to computer adaptive testing, where the computer programme selects the next item the candidate has to solve, based on the correctness of their answer to the previous item. If the candidate answers the item correctly, the next item will be a more difficult one, whereas if they answered the item incorrectly, the next one will be an easier one. In in this way, the main advantage of the test is that it adapts to the abilities of the test taker (Larson, 1989; Sulaiman & Kahn, 2019). However, such adaptive tests also have disadvantages. For instance, in contrast with linear tests, adaptive tests do not present the same set of items to each test taker. For this reason, the process of task development for such a test is more time consuming and it requires more financial investment than a linear test. A vast item pool is also important to ensure the security of the test items (Wainer & Eignor, 2000).

Using computers to interact with the candidates is another popular way to include the use of computers into the testing environment. One of the main advantages of such practice is that it enables the use of multimedia material in language tests, which can enhance the authenticity of the test, especially in the case of listening tests (Noijons, 1994). However, it also raises concerns because the richness of the visual input can be disturbing for the candidate (Noijons, 1994). In addition, if the candidates are not familiar with the digital platform, the new type of environment can cause stress or anxiety for them.

The last phase of language examinations where computers can be utilised is the phase of evaluation. According to Noijons (1994), in connection with evaluation, the most basic option is to use the computer to score the participants' answers. This can be especially well done in the case of multiple choice or true or false items. However, the evaluation can be problematic in the case of short answer items and writing tasks as the computer is not capable of making judgement in these cases without the assistance of a human assessor. In addition

to scoring, the computer can also provide meta-data about the task solving processes and strategies of the participants, and it can compare their results with the results of previous candidates (Noijons, 1994). Such data can be especially useful for research purposes.

In the case of the present dissertation, the computer platform was used only for administering and scoring the tests. The test tasks were pre-selected and every candidate of a certain language proficiency level received the same set of tasks, with the tasks and the items being in the same order, so the computer was not utilised in the generation of the test at all. In the present study, the main advantage of the computer platform was that it enabled the use of multimedia material in the test. Furthermore, it also automatically scored the candidates' answers. In the case of the short answer items, a list of acceptable items was programmed into the software; however, those answers, which the computer deemed incorrect, were also double-checked by a human assessor to ensure the reliability of the correction. Initially, the collection of meta-data with the help of the platform was also taken into consideration; however, this could not be accomplished by the time of the data collection.

Another test method facet besides the testing environment is the *test rubric*. Test rubrics (Bachman, 1990) include those factors which are connected to the organisation of the test (e.g., reading, listening, writing, and speaking components of the test), the time allocated to complete the different sections of the test, and the explicitness of the instructions of the tasks. Considering a computer-based listening test, for example, the explicitness of instructions is crucial because the unfamiliar context originating from the test environment (i.e., sitting in front of a computer and not in front of a piece of paper) might be counterbalanced by the user-friendly presentation of the information and the instructions on the screen (e.g., scrolling on the screen or dragging and dropping chunks of text from one part of the screen to another part of the screen to answer an item).

One of the most problematic characteristics of any kind of measurement is that one can only be certain about what is being measured (i.e., the reliability of the evidence) to an extent to which one is certain about the reliability of the measuring device itself. In hard sciences (e.g., chemistry and physics) this claim is an important approach to how measurement as such should be treated. When it comes to soft sciences (e.g., psychology and pedagogy), this characteristic feature of measurement should be treated even more carefully. In language testing, for example, the nature of the measured object, language competence, is a latent characteristic of the human mind (McNamara, 1996). Therefore, in order to make it observable to some extent, different tasks should be used. On the basis of such tasks it is possible to make inferences about test takers' language competence (Fulcher & Davidson, 2007). The tasks, or in other words, *the nature of the input the test taker receives* (Bachman, 1990), involves performance affecting factors at two levels. One of the levels is the format of the input, which refers to the communication channel, that is, the input can be aural, visual or it can contain both of these channels. This test method facet is also an important factor in the present study as participants had to solve audio-visual tasks. The other level is the language of the input which consists of, for example, the propositional content, the grammatical structure, and pragmatic features of the instructions, and the discourse and schematic structures of the recordings in case of a listening test.

The fourth facet in Bachman's (1990) taxonomy is the *nature of the expected response to the input*, which can be associated with item format, that is, what types of language should be used to complete the tasks. There are some item formats (e.g., multiple choice, true or false statements, and choosing the correct information from two options) which require less production on the part of the test-taker, while other item formats (e.g., short answers, open-ended questions, and fill-in-the-gaps tasks) involve more productive work by the test-taker. The nature of the language and the restrictions on response

(Bachman, 1990) – which are both components of this facet – are also connected to the item format because one particular item format entails certain grammatical structures in the response to the input.

The fifth facet in the Bachman's (1990) categorisation describes *the relationship between the input and the response*. When it comes to a speaking test, for example, the relationship between the input and the response is more straightforward to understand because there is oral interaction (i.e., "reciprocal relationship") between the test-taker and the examiner. In terms of a listening test, however, the reciprocity between the input and the response cannot be established; there is no immediate feedback given to the response (i.e., the relationship is non-reciprocal).

Investigating listening comprehension from the point of view of the input can further reinforce its complex nature. In testing listening comprehension skills, the use of another skill is always necessary (e.g., reading, writing, and speaking) in order to solve a listening task. Therefore, the relationship between the nature of the input and the response is always multi-faceted and generally non-reciprocal unless there is a speaking test where there is a reciprocal relationship between the aural (and visual) input and the oral response. Furthermore, listening task types always tap into the short-term memory of the test taker (Berg, 1987; Bregman, 1978; Buck, 2001). This feature of the listening tasks might become crucial, as far as speed and power listening tests are concerned. The notions of *speed* and *power* are connected to the test rubric facet discussed by Bachman (1990). While speed listening tests are based on the pace by which the information in the input occurs (e.g., a 100-item long sound-discrimination type of listening exercise), power listening tests are based on the quality of the input (e.g., answering multiple choice or open-ended questions about the gist of the text).

To counterbalance the possible difficulties emerging from the problems with short-term memory and input quality, in listening tests the recorded texts are usually repeated, which can be accounted for as another test rubric type of method facet. According to several researchers (Berne, 1995; Cervantes & Gainer, 1992; Iimura, 2007; Otsuka, 2004; Sakai, 2009), the repetition of the input entails better test scores. However, it also results in decreasing item discrimination values (Fortune, 2004), that is, the items cannot discriminate well enough between low-performing test takers and high-performing test takers. In other words, the possibility of repetition decreases the gap between the low- and high-performing test-takers, which might distort the reliability of the evidence regarding test scores. However, according to other studies (Field, 2009, 2013), it is more important to realise that test-takers use different levels of information processing in the two different attempts of listening; test-takers use lower-level processing in the first attempt and higher-level processing in the second attempt of listening (i.e., the opportunity of second listening allows students to concentrate more on checking and reformulating the discourse of their answer in order to produce a better response to the input).

As far as the input is concerned, it is also possible to take into consideration the test-taking strategies, which can be connected to both the nature of the input and the relationship between the input and the response according to Bachman's (1990) taxonomy. In terms of the nature of the input, there might be differences in the ways how test-takers answer questions if, for instance, the recordings are listened to first, then test-takers are presented with the series of questions (i.e., test management strategies are different in different contexts) (Cohen, 2006, 2011). If the language or the schematic context of the items are worded poorly, it is possible for the test-takers to answer questions without listening to or understanding the recording; in other words, test-wiseness strategies (Cohen, 2006, 2011) can also affect the relationship between the input and the response.

2.5 The concept of validity in language testing

Another key issue which should be considered in connection with language testing is *validity*. The ultimate aim of testing is to measure — through simulated tasks in the test — how well the candidate would perform in a real-life situation; in other words, what the relationship is between the test performance and the criterion performance. This relationship establishes the cognitive validity of the test (Glasser, 1991). The correspondence between the performance in real-life situations and the performance in the testing situation is referred to as the concept of *criterion-related validity* (Cohen, Manion & Morrison, 2000).

Regarding the concept of validity, however, it is also necessary to discuss the testing methods which became related to the constructs of the test itself (Bachman & Palmer, 1982). Separating testing methods and constructs, indeed, is a challenging task, but it has to be done because primarily it is the construct (i.e., language competence) which language testers are interested in, in every language testing situation. As far as epistemology is concerned, psychological traits like language competence cannot be touched or seen, and their presence and the extent of this presence can only be estimated through different tools. In language testing these tools are the methods, or to be more specific, they are the different tasks the candidates have to solve in the test. The only problem is that these tools affect the results, and they may lead to false interpretations about the presence and the degree of the traits.

The mainstream validation theory was born within the epistemological tradition and with Messick (1989), for whom the term *validity* is primarily determined by his notion of *construct validity* (i.e., whether the results of the test mirror what the test is meant to measure). Messick (1995) also argued that validity is not the characteristic of the test but that of the decisions which are made on the basis of test results (i.e., test scores). He also identified two types of threats to this validity. One of them is construct underrepresentation, which means that that test covers less than it should; namely, what candidates have to do in

real-life tasks are not represented well enough in the testing situation. The other type of threat he discusses is construct-irrelevant variance which means that the test measures something different from the construct, often alongside the construct itself; namely, the specificities candidates have to solve in the test are irrelevant from the construct point of view. With regard to construct-irrelevant variance, Messick (1995) distinguished between construct-irrelevant difficulty and construct-irrelevant easiness. He also proposed the question whether what makes a test easy or difficult is related to the construct or not, and whether this difficulty or easiness is due to some method effects or not. It follows from these questions that whoever is responsible for the test should be aware of different interpretations of the same test score and discount different rival interpretations.

Messick was criticised for not providing methods to carry out his theoretical validation framework. Kane (2004) following Messick's tradition provides an argument-based validation framework for dealing with such questions. His argument-based validation consists of two arguments: interpretive arguments (i.e., test scores and other evidence confirming or disapproving intended interpretations) and validity arguments (i.e., other possible interpretation, for example, by different stakeholders) with the help of which different interpretations can be verified or falsified.

A completely different approach was put forward by Borsboom, Mellenbergh, and van Heerden (2004), which contains a fundamentally different point of view compared to Messick's (1995) and Kane's (2004) mainstream validation tradition. Borsboom et al. (2004) argue that *validity* is the property of the test, and they made this claim on an ontological basis. While the mainstream epistemological validation is tasked with discounting rival interpretations, the ontological type of argument entails that validity should demonstrate purely causal relationship. That is, while Borsboom et al. (2004) argued for trusting in one's judgements and making immediate judgements on the basis of the test

results by creating a causal relationship between the attribute (i.e., task or item) and the reference (i.e., meaning), the mainstream epistemological validation tradition proposes not to trust one's immediate judgements and not to immediately connect something to the test because it may not be due to the test but to particular method effects. In terms of analysing test scores, this means that a correlation between two traits does not mean that one trait affects the other because correlation does not automatically create causality. In fact, that is the reason why the validation framework proposed by Borsboom et al. (2004) was not welcome in the language testing research community.

The present study, therefore, intends to follow the mainstream validation tradition. Following the mainstream validation approach is especially important since the construct validity and the criterion-related validity of current listening tests might be questioned, as present day practice suggests that engaging with multimedia and audio-visual material became part of people's everyday life both related to work and education (Brynjolfsson & Mcafee, 2014). For this reason, it can be presumed that the criterion-related validity of an audio-visual task must be higher than that of the audio-only task because it maximises the reflection of the real-life situation.

2.6 Using audio-visual materials in education

With the rapid advancement of technology, the regular consumption of audio-visual material has become part of people's daily life. The act of audio-visual reception can be defined as the following: "the user simultaneously receives an auditory and a visual input" (Council of Europe, 2001, p. 71), or "the user watches TV, video, or a film and uses multi-media, with or without subtitles and voiceovers" (Council of Europe, 2018, p. 54). In contrast with listening only activities, in case of audio-visual comprehension, the listener has to comprehend both audio and visual input.

Research, such as the one carried out by Greenberg and Zanetis (2012), suggests that currently education is going through a significant shift caused by the rapid spread of accessibility of diverse technological inventions. As mobile phones, laptops and access to the Internet became part of people's everyday life, these technological inventions started to also permeate the world of work and education. This opens new opportunities to enhance the quality of learning and teaching (Greenberg & Zanetis, 2012). To adapt to the changing social environment, videos have also been used in second language teaching for a long time, since the early 1980's (Rivers, 1981), because the non-verbal components are thought to be helpful in improving the listening comprehension skills of foreign language learners, especially in the case of beginner students (Carr & Duncan, 1987; Lonergan, 1984; Rubin, 1995). Research carried out in the field of language pedagogy has shown that using audio-visual or multimodal material in foreign language teaching can help students remember more details from the material than traditional audio-only tasks (Folley, 2015). Furthermore, Suvorov (2009) claims that using audio-visual material in foreign language teaching is useful for the language learner because the visuals can improve the authenticity of the communicative situation presented in the task, and they can help the listener identify the speaker roles. Being able to see the kinesic elements of the communication can also aid the listener in a more accurate understanding of the communication (Suvorov, 2009).

Furthermore, the Kaltura Report (2019) also supports the claim that using audio-visual material in education is an increasing trend. In their study conducted with educators, educational professionals, and students from all over the world, 82% of the participants asserted that using audio-visual materials should be part of the learning process, and 86% see helping students develop the necessary digital skills as a duty of the educators. The survey also suggests that using audio-visual materials in an interactive way can enhance the learning experience and increase the level and quality of the student achievement.

In addition, the study found that 95% of the respondents agreed that adequate digital skills are essential in finding a job and being successful at the workplace (Kaltura Report, 2019).

Digital literacy can be defined in several different ways, for example, Johnson, Adams Becker, Estrada, and Freeman (2015) define it as "being competent with a wide range of digital tools for varied educational purposes, or as indicator of having the ability to critically evaluate resources available on the web" (p. 24). This seems to be a rather narrow definition of the term related only to the domain of education; however, the Kaltura Report (2015) provides a broader definition which can be applied for all domains of life. According to this report (Kaltura Report, 2015), digital literacy is "the ability to locate, organise, understand, evaluate, analyse, create and communicate information using digital technologies" (p. 5). This definition shows that possessing adequate digital literacy skills is not only essential for being able to successfully participate in education but also a requirement for achieving success in one's field of work. In developing such skills, the Kaltura Report (2015) predicts that the role of the educator will considerably change, moving from the centre stage to the side lines, and becoming only the facilitator of the learning instead of the sole beholder of knowledge.

Audio-visual material used in education can be categorised in several different ways. Firstly, Hansch, Newman, Hillers, Shildhauer, McConachie and Schmidt (2015) categorised different types of videos based on nine features which they called the 'different affordances of Video'. The nine different types of videos are the following: (1) *building rapport*, which intends to create an emotional connection; (2) *virtual field trips*, which provide an opportunity to learn about people and places; (3) *manipulating time and space*, which gives access to new perspectives, such as slow motion or micro- and macro-views; (4) *telling stories*, which can captivate the viewers' attention and bring them along on a journey; (5) *motivating learners*, which can evoke the students desires for knowledge; (6) *historical*

footage, which reanimate past eras; (7) *demonstrations*, which can provide illustrations of experiments; (8) *visual juxtaposition*, which can help illustrate the contrast between concepts; and (9) *multimedia presentation*, which can utilise the combination of a wide variety of audio-visual elements.

Another categorisation is provided by the model of Siemens, Gašević and Dawson (2015) entitled 'The impact of networks on learning'. This model differentiates between videos directly created by the institution, the teacher or the students with an educational purpose in mind and videos publicly available on the Internet created by an outside party. In the case of the first type of video, the production distance between the educators and the used video is smaller as the educators themselves can partake in the design and the production of the video, and they can even appear as actors or as voiceovers in the video. In contrast, the latter type of video is usually not created for being specifically used as part of the course material, but it is selected by the tutor because it can successfully explain or illustrate a specific point or concept related to the course material. In using such sources, the educator should bear in mind respecting the relevant copyright laws (Siemens, Gašević & Dawson, 2015). Researchers such as Yousef, Chatti and Schroeder (2014) consider the use of videos created by institutions, educators or students the most useful in the educational context because in this way the videos can be specifically designed and customised for the didactic goals of the course. Audio-visual content generated by students can also successfully aid the students' self-reflection about their learning processes or they can be used as a form of assignment (Yousef, Chatti & Schroeder, 2014).

The most comprehensive taxonomy for classifying the different types of videos which can be used for educational purposes is provided by Woolfitt (2015). He offers an overview of the different types of videos by merging the previously available taxonomies and categorisations and distinguishes 16 different types of videos. Some examples of these

types are clips and fragments of YouTube videos, documentaries, Live Lecture Captures, Webinars, and Google HangOuts. For a full list of the video types and their detailed descriptions see Woolfitt (2015, pp. 18-20).

The above categorisations show that there is a wide variety of videos which can be used to aid education. However, research also suggests that because of the passive nature of consuming videos, their use should be executed with caution. De Boer (2013) claims that when considering using audio-visual materials in education, especially videos, the perspective of constructivism should be kept in mind. According to the constructivist theory (Simons & Bolhuis, 2004), students should be active participants in their learning process through constructing relevant knowledge. Relevant knowledge is constructed by linking new information with already existing knowledge and beliefs (Simons & Bolhuis, 2004). For this reason, when using videos in teaching, it must be ensured that the students' active participation in the learning process is stimulated and facilitated (De Boer, 2013).

Furthermore, the concept of *cognitive overload* must also be taken into consideration. Mayer and Moreno (2003) assert that in order to avoid cognitive overload, multimedia content used in education should create a balance between the visual and audio input, and only elements which foster the learning process should be employed. This claim suggests that designing multimedia aided course content should take into consideration the different elements of human understanding, and they should be designed strictly based on the learning goals of the course. Furthermore, these aspects should also be considered when designing multimedia aided tests to assess students' performance.

Using audio-visual and multimedia content in tests and in language tests in particular, is not widespread at the moment, especially in the Hungarian context. Nevertheless, the above discussion shows that such practice is proliferating in the field of education. For this reason, to ensure the maximization of criterion-related validity in

language tests, it can be proposed that it is time to reconsider the listening comprehension component and raise the possibility of supplementing it with the use of multimedia and audio-visual material. Therefore, the next section of the dissertation provides an overview of the research conducted in the field of testing with audio-visual material.

2.7 Testing audio-visual comprehension

Even though the use of audio-visual material in foreign language teaching is gaining popularity, test developers seem to be reluctant to include videos in language tests for measuring listening comprehension. Even the revised version of the CEFR devotes only one single scale to audio-visual comprehension (Council of Europe, 2018). The audio-visual reception scale focuses on three main concepts: the ability to understand and follow the main ideas, the ability to comprehend details and implied meaning, and the ability to understand different types of language use. Compared to the other competences described by the CEFR (Council of Europe, 2018), audio-visual reception appears to be heavily underrepresented.

The reluctance of test developers to include audio-visual material into language tests can have several explanations. Although researchers like Progosh (1996) and Wagner (2007) promote the use of audio-visual tasks in tests by claiming that in real life situations the non-verbal elements of communication are just as important as the verbal ones, one of the most often made criticism against using audio-visual tasks in language testing is that these tasks might measure something different from listening comprehension (Buck, 2001). The possible construct-irrelevant variance could be a relevant concern. However, it must not be forgotten that audio-visual tasks can emulate real-life situations, namely, target language use (Bachman, 1990), better than tasks which only involve audio input.

Numerous studies tried to investigate how L2 listening performance is influenced by using different types of audio-visual tasks, and their results are contradictory (Kellerman, 1990; Ockey, 2007; Raffler-Engel, 1980; Sueyoshi & Hardison, 2005). The different visuals

used in researching the role of visual input in listening comprehension can be divided into four categories: context related images (e.g., a still image depicting two people talking to each other on the street), content related images (e.g., the photo of a figure or a table accompanying a presentation about it), context related videos (e.g., a video recording of two people talking to each other in a classroom), and content related videos (e.g., the video recording of a set of presentation slides) (Suvorov, 2011). It can be presumed that the different types of visuals can have different effects on the test taker's performance. Even though this issue has already been investigated in the past, the results of the studies are not congruent with each other. The most notable studies in this topic are Bejar, Douglas, Jamieson, Nissan, and Turner (2000), Ginther (2002), and Ockey (2007). Bejar et al. (2000) investigated TOFEL test takers' performance when different types of visuals are included in the listening tests, and they found that including pictures which provide information about the context of the situation positively influenced the test performance of the candidates. Ginther (2002) arrived at similar results by finding that visual input which complemented the content of the aural input had a positive effect of the test takers' performance. However, she also found that context related visuals had a negative effect on understanding short talks, no significant effect on understanding conversations, and positive effect on understanding lectures. On the contrary, Ockey (2007) claims that in his study, several test takers reported that they made no use of the visuals, and that they were not even looking at them; therefore, the visual material had no effect on the participants' comprehension. Londe (2009) arrived at similar results. She created two different video recordings (i.e., a recording of only the presenter's face and a recording of the full body of the presenter) and an audio-only recording of the same 10-minute lecture, and they were used in a quasi-experimental research format. The participants were divided into three groups, and each group watched a

different recording. According to Londe's findings (2009), the performance of the participants was not influenced by either of the recordings.

The differences in the test takers' attitudes have also been investigated by other researchers, and they did not arrive at unequivocal results either (Dunkel, 1991; MacWilliam, 1986; Ockey, 2007; Sueyoshi & Hardison, 2005; Wagner, 2002). Researchers, such as Dunkel (1991), Sueyoshi and Hardison (2005), and Wagner (2002) all found that students preferred audio-visual tasks over the audio-only ones, and that the presence of visuals positively affected their test performance. In contrast, for example, MacWilliam (1986), and Ockey (2007) found that their participants claimed that they were not watching the video accompanying the audio input because they found it distracting. Although all these pieces of research are equally well designed, they arrived at contradictory results. Nevertheless, it should be considered that researchers like Kirschner and van Merriënboer (2013) suggest that students often are not able to adequately judge the efficiency of the methods they are employing, so their claims of preference and their attitudes towards using or avoiding audio-visual tasks should be examined with care.

The potential for being a distractor rather than a facilitator in tests is an often-raised concern in the debates about using audio-visual tasks in testing. However, the major difference between audio-visual and audio-only tasks is that in comparison with the traditional, audio-based listening task, in audio-visual tasks, the candidate can also rely on the speakers' kinesic behaviour (Raffler-Engel, 1980). Kinesic behaviour is a natural, non-redundant part of oral communication, which involves body language, facial expressions, gestures, and visible stress patterns (Kellerman, 1990; Raffler-Engel, 1980). Both Kellerman (1990) and Raffler-Engel (1980) argue that kinesic behaviour is natural, non-redundant part of verbal interaction because when there is a higher chance for misunderstanding, the speakers' kinesic behaviour increases. Moreover, when information

deduced from the linguistic and the kinesic input are contradictory to each other, listeners tend to accept information deduced from the kinesic input over the linguistic one (Burgoon, 1994). This fact further reinforces the presupposition that audio-visual tasks emulate real-life situations more closely than the traditional audio-only based listening tasks. However, as the results of previous studies are inconclusive and some of them lack enough questionnaire or interview data from the test-takers, further research is needed in the topic. The present study aims at contributing to the remedy of this research hiatus.

3 Research questions

The aim of the present dissertation is to analyse whether including audio-visual tasks into the listening comprehension component of language examinations is necessary and desirable. In order to do so, the study intends to answer the following research questions:

1. Do the paper-based sets of tasks and the computer-based sets of tasks measure listening comprehension in an equally reliable way?
2. Does the performance of the test-takers on the audio-visual-to-audio-only tasks differ from their performance on the audio-visual tasks?
3. Do the participants perceive the inclusion of audio-visual tasks as useful?

4 Research methods

As the proposed research aim promotes change on the Hungarian foreign language testing scene, the adopted philosophical background is the advocacy and participatory worldview (Creswell, 2009). This worldview intends to initiate change in a certain practice, and it promotes an open discussion about the issue in question. In the light of this, the present study analyses whether the listening comprehension component of foreign language examinations could be supplemented with audio-visual tasks, and it wishes to raise and discuss this possibility. To be able to do that, the possibility of using audio-visual material in language testing had to be examined at different language proficiency levels and from different perspectives. Therefore, the study adopted mostly quantitative strategies of inquiry to be able to describe trends and attitudes in a variety of samples, and it used language proficiency tests and questionnaires as its main data collection instruments. At the same time interviews were also used to obtain qualitative data about the participants' perceptions of the task in order to gain a better insight into their perspectives and to use these in the process of improving test methods.

The data collection procedure was conducted in three phases from the beginning of September 2017 to the end of August 2018. The majority of the data was collected in the framework of a larger language examination development project. The original aim of this language examination development project was to develop a computer-based language examination for four language proficiency levels (i.e., A2, B1, B2, and C1), and it was carried out by a major Hungarian language school. As the questionnaire development did not form part of the language examination development project, the data collections of the first and second phase were carried out by the author of this dissertation independently from the project. For this reason, all of the participants of the first and the second data collection phase were recruited by the author himself.

The following sections present the various steps of data collection and data analysis. For a more logically organised overview, the relevant information is organised along the lines of the three phases of the data collection (i.e., task development phase, questionnaire development phase, and pre-test phase).

4.1 Data collection and data analysis procedures

First and foremost, it is important to emphasise that the present study is part of a larger language examination development project undertaken by a major Hungarian language school. Therefore, most of the data collection was conducted in the framework of this language examination development project, and several decisions regarding the research design were directly and indirectly influenced by the official requirements of the Nyelvvizsgáztatási Akkreditációs Központ [Educational Authority Accreditation Centre for Foreign Language Examinations], (henceforward referred to as NYAK, following the Hungarian abbreviation) (Akkreditációs kézikönyv [Accreditation Manual], 2018]), which every accredited Hungarian language examination has to fulfil. In the framework of this project, an English and a German language examination were developed. In this way, the present study works with data obtained from tasks written in both English and German language. The source texts used for the tasks were not translations of each other or translations of texts written in other languages; they were authentic texts produced in English or German respectively. For the sake of a more transparent and easy-to-follow overview, the following section is organised along the lines of the steps of the data collection procedure.

4.1.1 First phase: Task development

The members of the language examination development team — assembled by the Hungarian language school responsible for the language examination development project — created 8 audio-visual (4 in English and 4 in German), 8 audio-visual-to-

audio-only (ATAO) tasks (4 in English and 4 in German) and 60 audio-only listening tasks altogether (30 in English and 30 in German). The term *ATAO tasks* refers to such tasks in which the visual material was removed from the originally audio-visual recording used as the source text for the task.

The task development team had nine members — one of them being the author of the present dissertation — and every member had at least three years of teaching experience, and many of them also had considerable experience in item writing. The tasks were developed for four different language proficiency levels, namely A2, B1, B2 and C1 levels. As the tasks were originally designed for the purpose of submitting them during the process of accreditation for the language examination, they had to be aligned with the CEFR descriptors (Council of Europe, 2001) referring to listening comprehension. Therefore, the first step of the task development was to study the listening comprehension in the CEFR (Council of Europe, 2001) descriptors for each language proficiency level. For a detailed overview of the CEFR (Council of Europe, 2001) descriptors and their requirements for each language proficiency level regarding listening comprehension and audio-visual reception, see section *4.1.1.2* (p. 51). In addition, at the beginning of the task development procedure, all members were required to complete a CEFR familiarization training in which they first became familiar with the CEFR scales (Council of Europe, 2001), then they had to pass a test on the descriptors.

As the second step of the task development, the 8 audio-visual, the 8 ATAO and the 60 audio-only English and German tasks were subjected to internal moderation. During the internal moderation, each member of the task development team was asked to revise and try out 2 tasks by themselves, written by another member of the team. Then, the tasks were finalised based on the feedback gathered during the moderation, and they were organised into sets of tasks intended for paper-based and for computer-based tests. The tasks used for

the paper-based and the computer-based test were different at each language proficiency level. This was necessary for two reasons: first, the paper-based test format was only intended for the pilot phase, to be able to compare the reliability of the computer-based tests to the paper-based tests; secondly, there was no time in the project for a test-retest format, and the data had to be collected with the same participants for the two versions. For a detailed overview of each set of tasks for each language proficiency level, see Appendix 1A-16A.

As the study also aimed at investigating the participants' perceptions of the finalised sets of tasks, a questionnaire was developed for this data collection purpose. However, no similar questionnaire was found in the literature so it had to be developed from scratch. Therefore, independently from the language examination development project, the author created an interview schedule which was intended to be used to collect data for the questionnaire development, and it was administered to the participants after solving one of the sets of tasks. For these interviews, the participants were recruited from various different language school groups taught by the author and fellow teachers. Participation in the interviews was done on a voluntary basis. The students were informed about the opportunity during class-time, and an appointment was fixed with those who were interested in participating.

For the data collection, each participant met the researcher individually on two different occasions. First, they solved the set of tasks intended for the paper-based test, and then the one intended for the computer-based test. At the end of each occasion, the participants were interviewed about their opinion and experiences regarding the test. Each data collection occasion took approximately 60 minutes (approximately 30 minutes for solving the tasks, and 30 minutes for the interview). Before the beginning of the data collection, the students were given the following information: (1) participation in the study happens on a voluntary basis; (2) they can withdraw from participation at any point of the

procedure, and withdrawing from participation would not have any consequences; and (3) for the protection of their personal data, their names are changed in the dissertation and any other publication resulting from the research study. The participants were also asked to sign a consent form (see Appendix 1E & 2E).

During the data collection occasions, the participants had to solve the tasks appropriate for their language proficiency levels. Their language proficiency levels were decided based on their placement test results administered by the language schools they attended. The interviews were conducted in Hungarian, the native language of the participants. For a detailed discussion of the interview schedule, see section *4.1.1.2* (p. 51). Because the digital platform used for administering the computer-based sets of tasks was the property of the language school responsible for the language examination development project, and this part of the data collection was carried out in other institutions, completely independently from the project, the participants filled in both test types on paper in printed format. In the case of the audio-visual task, the video was played on the researcher's laptop.

The data collected during the interviews was transcribed and subjected to content analysis. The categories emerging from the content analysis served as the cornerstones for the questionnaire developed during the second phase of the data collection procedure. However, as the circumstances of the data collection were different from the data collection circumstances of the third phase of the study, the test results of these participants were not taken into consideration when answering the research questions.

4.1.1.1 Participants of the first phase

Fifteen participants (9 males and 6 females) between the ages of 18-56 were involved in the interview studies leading up to the questionnaire development phase. They all participated in English and German language courses of different proficiency levels organised by different language schools in Hungary. The language proficiency levels of the

courses ranged from A2 to C1. The participants of the English level courses were all taught by the author of this dissertation himself; whereas the students of the German language courses were obtained with the help of a German language teaching colleague. The language proficiency level of the students was tested with the help of the different language schools' own English and German placement tests when they were placed into the most appropriate language course groups for their levels. Besides their language proficiency levels, the biographical data of the participants was also collected during the interviews. For a summary of this data, see Table 1 for the participants solving the English language tasks, and Table 2 for the participants solving the German language tasks. To protect the personal data and identity of the participants, they were given pseudonyms.

Table 1

Biographical Data of the Participants Solving the English Language Tasks in the First Phase

Name	Gender (she/he)	Language proficiency level	Age	Occupation	Years of learning English	Number of English classes per week at the language school	Other foreign languages	Future plans in connection with learning English
Liza	she	A2	56	shop assistant	1 year	180 minutes/week	Russian (learnt at elementary and high school)	moving abroad
Helga	she	A2	19	university student (1st year)	5 years	90 minutes/week	German (has C1 level language certificate)	B2 level language examination for university purposes
Bence	he	B1	19	high school student (13th grade)	5 years	270 minutes/week	French (learnt at high school)	B2 level language examination for university purposes
József	he	B1	19	high school student (13th grade)	5 years	270 minutes/week	French (learnt at high school)	B2 level language examination for university purposes
Anna	she	B2	23	receptionist	8 years	90 minutes/week	Spanish (has B2 level language examination)	B2 level language examination for possible promotion at work
Lilla	she	B2	21	university student (3rd year)	6 years	180 minutes/week	German (learnt at elementary and high school)	B2 level language examination for university purposes
Zénó	he	B2	18	high school student (12th grade)	4 years	180 minutes/week	German (learnt at elementary and high school), Spanish (learnt at high school)	B2 level language examination for university purposes
Peti	he	C1	26	PhD student	15 years	90 minutes/week	Italian (learnt at high school)	C1 level language examination for university purposes

Table 2

Biographical Data of the Participants Solving the German Language Tasks in the First Phase

Name	Gender (she/he)	Language proficiency level	Age	Occupation	Years of learning German	Number of German classes per week at the language school	Other foreign languages	Future plans in connection with learning German
András	he	A2	19	university student (1st year)	1 year	180 minutes/week	English (has B2 level language certificate)	B2 level language examination for university purposes
Juci	she	B1	20	university student (1st year)	5 years	90 minutes/week	English (learnt at elementary and high school)	B2 level language examination for university purposes
Béla	he	B1	43	lathe man	2 years	180 minutes/week	Serbian (learnt at elementary and high school)	moving to Germany at the end of the year
Csilla	she	B2	50	entrepreneur	6 years	90 minutes/week	Russian (learnt at high school)	B2 level language examination for work purposes
Géza	he	B2	33	accountant	3 years	180 minutes/week	English (learnt at elementary and high school)	work purposes
Soma	he	B2	27	tourist guide	2 years	90 minutes/week	English (has C1 level language certificate) and Spanish (has C1 level language certificate)	work purposes
Ádám	he	C1	36	graphic designer	21 years	90 minutes/week	English (has C1 level language certificate) and French (has B2 level language certificate)	C1 level language examination for possible work abroad

The main aim of the participant selection was to find at least two participants for each proficiency level. However, this seemed to be difficult in the case of the A2 and the C1 language proficiency levels, as such courses seem to be less frequently requested in language schools. As the data presented in the table shows, the participants were between the ages of 18-56 and they had various language learning backgrounds. As most of them were preparing for various levels of language examinations, they proved to be ideal candidates for the interviews.

4.1.1.2 Data collection instruments of the first phase

First and foremost, the students of the language schools participating in the interview studies were selected on the basis of their results on a placement test administered by the language schools they were studying at. As the participants came from three different language schools, the placement tests were slightly different in each school. However, all the placement tests both in English and in German had two main parts: a written and an oral section. The first part contained test items targeting grammatical knowledge, where the items had a gradually increasing difficulty. The second part contained an approximately 10-minute-long oral discussion about general topics, such as family, studies, work or hobbies. Based on the results of the placement tests, the students were placed into the most appropriate course for their respective language proficiency levels. For confidentiality reasons, the exact tasks and details of the placement tests cannot be included in this dissertation.

The second data collection instrument contained 8 audio-visual tasks, 8 ATAO tasks, and 60 audio-only tasks organised into 8 sets of tasks intended for a paper-based test (4 English and 4 German), and 8 sets of tasks intended for a computer-based test (4 English and 4 German). Both the English and the German paper-based tests contained 3 audio-only tasks and 1 ATAO task on the A2 language proficiency level, and 4 audio-only tasks and 1

ATAO task on the B1-C1 levels. In the case of the computer-based tests, both the English and the German tests contained 3 audio-only tasks and 1 audio-visual task on the A2 level, and 4 audio-only tasks and 1 audio-visual task on the B1-C1 levels. For more details on the tasks, see Appendix 1A-16A. The terms *paper-based test* and *computer-based test* refer to the difference in the test delivery process. In the paper-based test, the participants had to sit the test in a paper-and-pen format; whereas in the computer-based test, they had to solve and answer the tasks on a computer.

As these tasks were part of a larger language examination development project, they had to be fitted to the respective scales and descriptors of the CEFR (Council of Europe, 2001). By the time of writing up this research study, a newer version of CEFR has already been published in 2018 (Council of Europe, 2018); however, as the task development and the first phase of the data collection for this study were conducted in the autumn of 2017, the present study used the 2001 version of the CEFR (Council of Europe, 2001). According to the CEFR (Council of Europe, 2001), the *Overall Listening Comprehension* scale defines the listening comprehension requirements for each language proficiency level as follows:

Table 3

Overall Listening Comprehension Scale (Council of Europe, 2001, p. 66)

Proficiency level	Descriptor
C2	Has no difficulty in understanding any kind of spoken language, whether live or broadcast, delivered at fast native speed.
C1	Can understand enough to follow extended speech on abstract and complex topics beyond his/her own field, though he/she may need to confirm occasional details, especially if the accent is unfamiliar. Can recognise a wide range of idiomatic expressions and colloquialisms, appreciating register shifts. Can follow extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly.
B2	Can understand standard spoken language, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal, social, academic or vocational life. Only extreme background noise, inadequate discourse structure and/or idiomatic usage influences the ability to understand.
	Can understand the main ideas of propositionally and linguistically complex speech on both concrete and abstract topics delivered in a standard dialect, including technical discussions in his/her field of specialisation. Can follow extended speech and complex lines of argument provided the topic is reasonably familiar, and the direction of the talk is sign-posted by explicit markers.
B1	Can understand straightforward factual information about common everyday or job related topics, identifying both general messages and specific details, provided speech is clearly articulated in a generally familiar accent.
	Can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure etc., including short narratives.
A2	Can understand enough to be able to meet needs of a concrete type provided speech is clearly and slowly articulated.
	Can understand phrases and expressions related to areas of most immediate priority (e.g. very basic personal and family information, shopping, local geography, employment) provided speech is clearly and slowly articulated.
A1	Can follow speech which is very slow and carefully articulated, with long pauses for him/her to assimilate meaning.

Following the suggestions of the CEFR (Council of Europe, 2001), the texts used for the development of the audio-only tasks were all based on authentic texts produced in real-life situations with a genuine communicative intention by native speakers of English and German. However, most often the audio quality of the original recordings was not satisfactory for language testing purposes because they originated from sources such as radio and TV reports and interviews. Therefore, they were re-recorded in a studio with the help of native speakers of English and German. During the selection of the texts it was also ensured that they had a clear structure with appropriate vocabulary and syntax for the particular

language proficiency levels, and that they were discussing a topic which prospective examinees above the age of 14 would be familiar with. The task developers also paid attention that none of the topics discussed in the tasks were offensive or upsetting in any way for the examinees. For this reason, the main topics of the tasks were related to everyday topics, such as lifestyle, relationships, environment, work, school, hobbies, economy, travelling, science and technology.

The listening activities presented in the tasks were also varied. The tasks contained situations such as listening to public announcements, listening to TV and radio recordings, listening to public events, and listening to conversations between native speakers. These listening activities were also chosen on the basis of the recommendations of the CEFR (Council of Europe, 2001), and they were calibrated according to the language proficiency requirements described in the listening comprehension sub-scales entitled *Understanding Conversations Between Native Speakers*, *Listening as a Member of a Live Audience*, *Listening to Announcements and Instructions*, and *Listening to Audio Media and Recordings*. Features such as the length of the text, the accent and the number of the speakers, and the speed of the text were also considered. For the guidelines that were followed regarding these features, see Table 4.

Table 4

Guidelines for Task Development

Performance factors	A2	B1	B2	C1
Accent of the speaker	Standard (American or British English; German – Hochdeutsch).	Standard.	Standard.	Standard, or slightly non-standard.
Number of speakers	Only two.	Two, maximum three speakers. The voice of the speakers has to be distinguishable.	Two, maximum three speakers. The voice of the speakers has to be distinguishable.	Two or more.
Acoustic characteristics of recording	Clear, without any background noise.	Clear, without any background noise.	Clear, with some background noise.	Clear or with genuine, i.e. realistic background noise.
Speed of speech and articulation	Clear, slow and well structured.	Relatively slow. Clearly articulated.	Standard and well-articulated.	Standard and fast.

The item formats chosen for the different tasks were also influenced by the intended language proficiency levels and by a review of the best practices from already operating language examinations (e.g., Cambridge, Goethe, TELC, and IELTS). The fact that the language examination development project intended to develop tasks for a computer-based interface also greatly affected the choice of item formats. Based on these considerations, as Table 5 shows, the following item formats were chosen for the different language proficiency levels:

Table 5

Task Types Used in the Research Project

Proficiency levels	Task types
A2	True or false Multiple choice (3 options) Fill-in the gap
B1	Multiple choice (3 options) Short answer Matching Fill-in the gap
B2	Multiple choice (3 options) Short answer Matching Fill-in the gap
C1	Short answer Matching Fill-in the gap

For a detailed and tabulated overview of all the main features of the test tasks administered during the data collection for this dissertation, see Appendix 1A-16A.

The audio-visual tasks and the ATAO were developed in a similar fashion to the audio-only tasks. The 'Watching TV and Film' scale of the CEFR (Council of Europe, 2001) served as the basis of the audio-visual and ATAO task development. According to the CEFR (Council of Europe, 2001), as Table 6 demonstrates, the audio-visual skills required for each proficiency level are the following:

Table 6

Watching TV and Film Scale (Council of Europe, 2001, p. 71)

Proficiency level	Descriptor
C2	As C1.
C1	Can follow films employing a considerable degree of slang and idiomatic usage.
B2	Can understand most TV news and current affairs programmes. Can understand documentaries, live interviews, talk shows, plays and the majority of films in standard dialect.
B1	Can understand a large part of many TV programmes on topics of personal interest such as interviews, short lectures, and news reports when the delivery is relatively slow and clear.
	Can follow many films in which visuals and action carry much of the storyline, and which are delivered clearly in straightforward language. Can catch the main points in TV programmes on familiar topics when the delivery is relatively slow and clear.
A2	Can identify the main point of TV news items reporting events, accidents etc. where the visual supports the commentary.
	Can follow changes of topic of factual TV news items, and form an idea of the main content.
A1	No descriptor available.

Similarly to the audio-only tasks, the videos selected for the development of the audio-visual and ATAO tasks were all authentic material produced by native speakers of English in real-life situations with real communicative intentions. The language of the video material was selected according to the vocabulary and syntax appropriate for the relevant language proficiency levels, and the topics discussed in the videos were topics which examinees over the age of 14 could easily identify with. Offensive or inappropriate content was also avoided. Therefore, the main topic areas covered in the audio-visual and ATAO tasks was the same as in the audio-only tasks.

Similarly to the audio-only tasks, length, accent of the speakers, speed of articulation and acoustic quality were also considered. The decisions regarding these features were also governed by the same specifications as in the case of the audio-only tasks. Furthermore, the item formats were also chosen in a similar fashion. For a detailed and tabulated overview of all the main features of the test tasks administered during the data collection for this dissertation, see Appendix 1A-16A.

The last data collection instruments in this phase of the research study were two semi-structured interview schedules. The semi-structured design was selected on the basis of the recommendations of Maykut and Morehouse (2002) because the freedom and flexibility ensured by this design appeared to be the most fitting for the purposes of the study. The interview schedules were created by following McCracken's (1988) four-step model for designing and implementing a long qualitative interview. The four steps of McCracken's model (1988) are the following:

(1) review of analytic categories and interview design

(2) review of cultural categories and interview design

(3) interview procedure and the discovery of cultural categories

(4) interview analysis and the discovery of analytical categories (p. 29).

The first drafts of the interview schedules were created by following the first two steps of McCracken's model (1988). After reviewing the literature on the topic of researching listening comprehension, a German language study (Porsch, Grotjahn & Tesch, 2010) was found to be useful for the present research study. The aim of Porsch et al.'s study (2010) was to examine the extent to which listening comprehension is influenced by visual input, and whether listening comprehension and audio-visual comprehension are the same construct. The study was conducted in Germany, and their sample was composed of 156 high school students (9th grade) whose first language was German and who studied French as a second language. The participants had to solve audio-visual and audio-only listening comprehension tasks, and then they had to answer questions about different aspects of the items. Even though they did not publish the full list of the questions used in the study, the paper gives a detailed and thorough description of the categories examined. These categories and the own self-reflections on the topic by the author of this dissertation served as the bases of the first drafts for creating the semi-structured interview schedules. After each interview,

the participants were asked if they had any further remarks on issues they had not been asked about, and based on these remarks, the interview schedules were continually improved. As the sets of tasks used for the two data collection occasions differed regarding the presence or the absence of the audio-visual task, two slightly different interview schedules were developed for the paper-based and for the computer-based tests. For the final versions of the semi-structured interview schedules, see Appendix 2B and 3B.

As the paper-based test version was administered first with each participant, the corresponding interview schedule had two main parts: questions related to the participants' biographical data and questions related to the listening comprehension tasks. Since the computer-based test version was administered during the second data collection occasion, the corresponding interview schedule only contained questions related to the audio-only and the audio-visual text comprehension tasks, and no questions about the biographical information of the participant. During the content analysis of the interviews, five main constructs seemed to emerge: *disturbing factors*, *the difficulty of the tasks*, *the quality of the tasks*, *the number of tasks in relation to the time given*, and *the helpfulness of the video*.

4.1.1.3 Analysis of the data collected in the first phase

As the first step of the data analysis, the interviews conducted with the participants were transcribed and subjected to content analysis. Based on the first interview, a coding scheme was developed, which was continuously expanded and improved as further interviews were analysed. To ensure the reliability of the coding, two interviews were co-coded by a colleague who is familiar with the methods of interview analysis. The Cohen's Kappa was calculated with the help of SPSS 22.0 as a measure of the inter-coder reliability. As the result was Kappa = 0.78 ($p < 0.001$), the coding was deemed reliable, and the emerging themes were identified.

Five emerging themes were identified: *disturbing factors*, *the difficulty of the tasks*, *the quality of the tasks*, *the number of tasks in relation to the time given*, and *the helpfulness of the video*. *Disturbing factors* can be defined as the unknown words and strange accents with which the participant has to cope with during the process of solving the task. The speakers' intonation in the recordings could possibly also be perceived as annoying, and it can also be part of this category. The same speaker's intonation might not be annoying to all the test-takers, but even with the best intentions on the part of a test design team, there might be speakers whose voice and intonation can cause a certain level of frustration in some test-takers. Furthermore, it is important to distinguish between the *difficulty of the tasks* and *the quality of the tasks*. The former one refers to the cognitive difficulty of solving a particular item, which means that the item might be too difficult for the test-taker in terms of its content, and not because of its format. The latter one refers to the extent to which test-takers find a particular recording interesting, thought-provoking or simply enjoyable. *The number of tasks in relation to the time given* category refers to the amount of time available to solve the tasks. This category also seems to be important since time-constraints can put extra unnecessary anxiety on test-takers in the testing situation. The last category, namely, *the helpfulness of the video* refers to the extent to which a particular audio-visual material helps test-takers to understand a particular recording better, or the extent to which the video is counterproductive and serves as a distraction from concentrating on what is being said in a particular audio-visual recording.

4.1.2 Second phase: Questionnaire development

During the second phase of the data collection, two questionnaires were developed from the data collected with the help of the semi-structured interviews. As it has already been mentioned, five major themes emerged from the interview data. Based on the five

themes and the constructs used in the study conducted by Porsch et al. (2010), two questionnaires were developed. The aim of the questionnaires was to gather feedback from the participants solving the tasks.

As one of the aims of the present dissertation was to compare the scores participants obtained on the paper-based and on the computer-based sets of tasks, two slightly different questionnaires had to be designed. The first questionnaire contained statements about issues related to executing the paper-based sets of tasks, whereas the statements in the second one referred to the computer-based sets of tasks. For a detailed discussion about the questionnaires, see section *4.1.2.2* (p. 63).

After the first versions of the questionnaires were developed, they were piloted with four English learners of different proficiency levels. During the pilot, the participants met the researcher individually on two different occasions. On the first occasion, they were asked to solve the paper-based set of tasks. When they had finished the tasks, they received a printed version of the corresponding questionnaire, and they were asked to indicate their agreement with statements about the tasks on a five-point Likert-scale while verbalising their thoughts about the items and their answers in a think-aloud format. The participants received the instructions for the think-aloud protocol in Hungarian (i.e., their mother tongue), but they were encouraged to execute the think-aloud in the language they felt the most comfortable with. This resulted in the participants using a mix of Hungarian and English language during the think-aloud protocols. The second pilot occasion happened in a similar way with the exception that the participants had to solve the tasks and had to fill in the questionnaire related to the computer-based sets of tasks. In order to familiarise the participants with the think-aloud method, a practice opportunity was provided for them in the form of a practice think-aloud task based on the recommendations of Bowles (2010) at the beginning of each data collection occasion. For the think-aloud practice tasks, see

Appendix 1B. For the reasons already mentioned in the discussion of the previous phase, on both occasions the participants received the test in a printed format, and they listened to the audio recordings and watched the video material on the laptop of the author of this dissertation. For this reason, the test results of these participants were also not taken into consideration when answering the research questions, and their answers for the questionnaire items were also only used to finalise the questionnaires.

On both data collection occasions, the participants' think-aloud protocols were audio recorded and later transcribed for content analysis. Finally, the questionnaires were finalised based on the feedback collected during this pilot. As the aim of the data collection in the second phase was not to validate the tasks but to validate the questionnaires, students of German were not involved because of the difficulty of access.

4.1.2.1 Participants of the second phase

The questionnaire was piloted with the help of four learners of English (one female and three males, their ages ranging from 21 to 36). These participants were all private students of the researcher and they spoke English at different language proficiency levels. Before the data collection, the participants were informed that participation in the study is done on a voluntary basis, and that to preserve their anonymity, their names are changed in the dissertation. They were also notified that they can withdraw from participation at any point during the data collection without any consequences, and they were asked to sign a form of consent (Appendix 1E & 2E) as an acknowledgement of this information.

During the first data collection, the participants were asked to do the paper-based test version and at the end fill in the corresponding questionnaire. During the second data collection occasion, they did the computer-based set of tasks and answered the corresponding questionnaire. As they were all private students of the author of this dissertation, their language proficiency levels were decided based on the author's intuition

and experience with them. Before the beginning of the data collection procedure, the biographical data of the participants was recorded (see Table 7).

Table 7

The Biographical Data of the Participants in the Second Phase

	Nóra	**Aladár**	**Zsombor**	**Márk**
Age	21	36	34	33
Language proficiency level	A2	B1	B2	C1
Years of learning English	3 years	5 years	7 years	11 years
Occupation	office worker	computer engineer	computer engineer	accountant

As the table shows, the participants were all adult learners who had already been studying English for a while. They all had different occupations and were preparing for different language examinations.

4.1.2.2 Data collection instruments of the second phase

Four main data collection instruments were used during the second data collection phase: the paper-based set of tasks, the computer-based set of tasks, and the two questionnaires. However, the following section only discusses the questionnaires. For a detailed overview of the tests, see Appendix 1A-16A.

The two questionnaires used in the second data collection phase had 28 items each, and they were administered in Hungarian, the native language of the participants. The items of the questionnaires were organised into five different constructs. The labels of the constructs were borrowed from the labels given to the emerging themes in the first phase. Therefore, the five constructs were *disturbing factors*, *the difficulty of the tasks*, *the quality of the tasks*, *the number of tasks in relation to the time given*, and *the helpfulness of the video*. The construct labelled *disturbing factors* contained items referring to possible difficulties (i.e., unknown words, background noise, or strange accents of the speakers) participants had to overcome while solving the tasks. An example of an item belonging to

this category would be the following: “The background noise on the recordings was disturbing”. The construct named *the difficulty of the tasks* contained items about the perceived cognitive difficulty of solving the items of the tasks. For instance, “I had to remember a lot of information at the same time to be able to answer the questions in the tasks”. *The quality of the tasks* contained items referring to whether the participant found the tasks enjoyable or thought provoking. For example, “The topics of the tasks were interesting for me”. *The number of tasks in relation to the time given construct* contained items such as “The time given for the tasks was enough”, and it referred to technical data about the length and the number of the tasks.

Up to the first four constructs, the questionnaires relating to both the paper-based set of tasks and the computer-based set of tasks were completely identical. The only difference between the two questionnaires was introduced in the items of the last construct, *the helpfulness of the video*. In the questionnaire concerning the computer-based set of tasks, this construct contained items asking about the perceived usefulness of the video accompanying the tasks. For instance, “The visual information in the videos helped my understanding”. On the other hand, in the questionnaire about the paper-based set of tasks, *the helpfulness of the video* construct contained items about the possible usefulness of adding a video to the recordings in the tasks. For example, “A video could have helped me answer the questions in the tasks”.

This first versions of the questionnaires were constructed on the basis of the semi-structured interviews conducted in the first data collection phase and based on the research conducted by Porsch et al. (2010). However, as during the second phase the questionnaires were only piloted with four participants, only the wording of the items could be finalised, but the reliability of the constructs could not be examined. Therefore, these 28-item versions are not considered to be the final versions of the two questionnaires

developed for this dissertation, and they are not added to the appendices. For a discussion of the finalised version of the questionnaires on the audio-visual and audio-only tasks, see section *4.1.3.2* (p. 68).

4.1.2.3 Analysis of the data collected in the second phase

As part of piloting the first draft of the questionnaires, four English learners were asked to solve audio-visual and audio-only tasks, and after that to fill in the two questionnaires while verbalising their emerging thoughts regarding the tasks and the questionnaire items (i.e., to perform a think-aloud protocol while filling in the questionnaires). The think-aloud protocols were audio recorded and transcribed after the data collection. Then the transcripts were subjected to content analysis and, based on the feedback of the participants, both the questionnaire referring to the audio-visual tasks and the questionnaire referring to the audio-only tasks were modified.

4.1.3 Third phase: Conducting the pre-tests

The original aim of the third data collection phase was to pilot the sets of tasks for the language examination development project. Therefore, the participating institutions were contacted by the head of the project, and the technical details of the organisation of the pre-test were also taken care of by the language school responsible for the project. During the pre-tests, the participating students solved the tests in class time in the presence of their language teachers. Solving a set of tasks took approximately 40 minutes so the paper-based and the computer-based sets of tasks were executed on two separate occasions. Therefore, in most participating institutions, two to three weeks passed between the administering of the paper-based and the computer-based sets of tasks.

At the data collection with the paper-based set of tasks, the participants received the full test in a printed format at the beginning of the data collection. The recording for each task was played twice from a CD before moving to the next task. The pauses allowed for

answering the items were also recorded on the CD. In the case of high schools and elementary schools, the computer-based tests were administered on the school computers in the computer laboratory. In the case of the language schools and the university, laptops were provided by the language examination development project for the same purpose. In the computer-based tests, the tasks had to be solved on a digital platform specifically designed for this purpose. This software was pre-installed on the laptops, and sent to the schools the day before the data collection. Both on the laptops and on the school computers it was ensured that the participants could not open and use any other software during the examination. The digital platform was programmed to provide approximately 35-40 minutes — depending on the language proficiency level — to complete the full listening comprehension component. Except for this time limit, there were no other individual time limits specified for each task. The participants could decide when to play and re-play the recordings. Nevertheless, the software was programmed to allow only two listening opportunities for each recording, and once a student moved on from a task to the next one, they could not return, even if they only listened to the previous recording once. In the case of the audio-visual task, if the participant decided to watch the video without looking at the task items, the video appeared in a larger window. However, the window of the video automatically shrank if the test-taker was scrolling down to the task items. This format was chosen as the best feasible solution for presenting the video and the items at the same time.

On both occasions, after solving the listening comprehension component, the participants received the questionnaire corresponding to the test format, and they were asked to give their honest feedback about the tasks. In the case of both test types, the questionnaire was provided for the participants in a printed format. Filling in the questionnaire took approximately 10-15 minutes, making the total time of completing the listening comprehension component approximately an hour.

4.1.3.1 Participants of the third phase

Altogether 140 students (60 males and 80 females) participated in the third phase of the data collection. They were between the ages of 12 to 42, and they came from several different contexts. The data collection took place at a major Hungarian university, in several groups of two major Hungarian language schools, in 6 high schools from 3 Hungarian counties, and 2 elementary schools from 2 Hungarian counties. The reason behind the selection of the participating groups and institutions was to obtain data from as many different institutions as possible. Moreover, involving only elite schools from Budapest, or only language school groups would have probably produced skewed results. The language proficiency level appropriate for the participants was decided by their language teachers and based on the coursebooks they were learning from.

The pre-tests were organised by the language school responsible for the language examination development project, and the institutions were also contacted by them. Originally more than 140 students participated in the pre-tests; however, those who did not execute both the computer-based and the paper-based tests were excluded from the present study. Thus, out of the 140 students, 73 executed the English tests and 67 participants the German tests. For the number of tests solved for each language proficiency level in each language, see Table 8 and Table 9.

Table 8

The Number of Participants Solving the English Tasks in the Third Phase

Language proficiency level	**Number of participants**
A2	11
B1	19
B2	26
C1	17

Table 9

The Number of Participants Solving the German Tasks in the Third Phase

Language proficiency level	Number of participants
A2	11
B1	19
B2	24
C1	13

4.1.3.2 Data collection instruments of the third phase

The first main data collection instruments of the third data collection phase were the paper-based and the computer-based sets of tasks developed during the first phase. For a summary of the main topics and other main characteristics of the paper-based and the computer-based sets of tasks designed for each language proficiency level, see Appendix 1A-16A. For the summary of the number of tasks in the different tests for each language proficiency level, see Table 10 and Table 11.

Table 10

The Number of Tasks in the English Tests

Language proficiency level	Paper-based test	Computer-based test
A2	4	4
B1	5	5
B2	5	5
C1	5	5

Table 11

The Number of Tasks in the German Tests

Language proficiency level	Paper-based test	Computer-based test
A2	4	4
B1	5	5
B2	5	5
C1	5	5

As the aim of the present dissertation was to investigate whether the listening comprehension component of foreign language tests should be supplemented with audio-visual tasks, and it does not intend to propose the inclusion of the audio-visual tasks as a separate component, but as part of the already existing listening component, the

reliability of the audio-visual tasks had to be examined as part of a listening comprehension set of tasks. This is the reason why no separate audio-visual set of tasks was developed. Furthermore, in order to gain a deeper insight into the usefulness and the necessity of the audio-visual tasks, in both the paper-based and the computer-based tests, the last task was a task originally created from audio-visual material. In the paper-based tests, the recording of the last task was modified by removing the visual material from the originally audio-visual recording. In this dissertation the term *audio-visual-to-audio-only* (ATAO) task is used to refer to such tasks. The removal of the visual material was necessary because of feasibility reasons as during the administration of the paper-based tests only CD-players were available for playing the recordings. In the computer-based tests, however, the test-takers had the opportunity to watch audio-visual material on the digital platform so the visual material of the last task could be played. The comparison of the participants' results on these two different tasks was important for answering the second research question because it could shed light on the extent to which their performance might have been influenced by the presence of the visual material.

The second main data collection instruments of the third phase were the two questionnaires developed during the first two phases of the data collection. At the beginning of the data collection of the third phase, the original 28-item versions of the questionnaires were administered among the participants. In order to be able to collect the background data of the students, three extra questions were added to the end of the questionnaires. The first one referred to the gender of the participant, whereas the second and the third questions referred to the level and type of task the participant had solved.

After the first 90 questionnaires were filled in, the data was entered in SPSS 22.0, and the Cronbach's alpha values of the constructs were calculated. As the initial constructs did not have a high enough internal consistency, 10 items were deleted from each

questionnaire, the remaining items were reorganised into four constructs, and these finalised 18-item versions of the questionnaires were used during the rest of the data collection. For a more detailed discussion about the process of finalising the questionnaire, see section *4.1.3.3* below. The final version of the two questionnaires can be found in Appendix 1C-2D.

4.1.3.3 Analysis of the data collected in the third phase

As part of the data analysis, the questionnaire data collected with the first 90 questionnaires was entered into SPSS 22.0., and the Cronbach's alpha values of the constructs were calculated. As the initial Cronbach's alpha values were dissatisfactory, the questionnaire items were subjected to a Two Step Cluster analysis. Based on the results of the cluster analysis, 10 questionnaire items were removed from each questionnaire. In both questionnaires, the following items were re-grouped into the following constructs: *disturbing features* (q2, q5, q7, q11), *structure of the test* (q1, q3, q6, q9, q10), *perceived difficulty* (q4, q8, q12, q14, q16), and *necessity of the video* (q13, q15, q17, q18). The *necessity of the video* construct has to be interpreted slightly differently in the case of the two different test formats: in the computer-based test questionnaire, it refers to the degree to which the participants found the video material useful for solving the last task; whereas in the paper-based test questionnaire, it refers to the degree to which the participants think some videos could have successfully aided them in solving the tasks. The Cronbach's alpha values of the new constructs show that the internal consistency of each construct is either acceptable or good (see Table 12 and Table 13). For the finalised versions of the two questionnaires, see Appendix 1C-2D.

Table 12

The Cronbach's Alpha Values of the Paper-Based Test Questionnaire Constructs

Name of the construct	Cronbach's alpha value
Disturbing features	0.70
Structure of the test	0.72
Perceived difficulty	0.79
Necessity of the video	0.85

Table 13

The Cronbach's Alpha Values of the Computer-Based Test Questionnaire Constructs

Name of the construct	Cronbach's alpha value
Disturbing features	0.72
Structure of the test	0.74
Perceived difficulty	0.71
Necessity of the video	0.83

For the rest of the data collection, these finalised versions of the questionnaires were used. At the end of the data collection, all the collected questionnaire data was entered into SPSS 22.0, and the mean values were calculated for each construct regarding both questionnaires, and ANOVA calculations were carried out for the *necessity of the video* construct. For a detailed report on the results, see sections *5.1.1* (p. 76), *5.1.2* (p. 103), and *5.1.3* (p. 107).

The test results were also entered in SPSS 22.0, and the Cronbach's alpha values of the tests and the variance of the test scores were calculated. Furthermore, the test results were entered in a Microsoft Excel spreadsheet to calculate the standard error of measurement (SEM) in the tests and the item facility values and point-biserial correlations of the items. As the results of these calculations were used to answer the first and the second research questions, the results of the analyses are reported in detail in sections *5.1.1* (p. 76) and *5.1.2* (p. 103).

The data analysis followed the classical test theory approach. This approach was chosen over the modern test theory approaches (e.g., item response theory) for several reasons. There are three different IRT probabilistic models which could be considered in

large-scale testing: the one parameter, the two parameter, and the three parameter model. The one parameter model is traditionally used for placing the test items on a difficulty continuum (Rasch, 1960), which was outside the scope of the present study. The two and three parameter models are designed to be used with samples with at least 1,000-2,000 participants or above — depending on the context — in order to produce accurate estimations. In such cases, the smaller the sample, the less reliable the results of item response theory approach are (Embretson & Reise, 2000; Hambleton 1989; Meunier, 1994). On the basis of this information, the present sample is too fragmented and too small for reliable IRT analyses.

4.2 Ethical considerations

As any research project involving the participation of human subjects, the present study also raises several ethical considerations. First, the principle of non-maleficence (Cohen, Manion & Morrison, 2000) should be taken into consideration. In order not to cause any harm or distress to the participants, in every phase of the study before beginning the data collection, the participants were informed about the details of the data collection procedure and about the ways their data is handled and stored during and after the research project. In addition, the participants of the first and the second phase were also provided with a consent form describing these issues in detail. For an example of the consent form used, see Appendix 1E and 2E. During the third data collection phase, no such consent form was used because the data collection occasions were organised by the head of the language examination project, and therefore, the participants were informed about the details of the data collection and data handling by the representatives of the project.

Second, based on the principle of beneficence (Kubanyiova, 2015), the participants should also gain some benefits from taking part in the data collection. In the present study, this benefit was the opportunity to practice language examination tasks and to get acquainted

with a new type of language examination carried out on a digital platform. As the majority of the participants were planning to take a language examination in the near future, the additional practice opportunity was probably valuable for them.

Third, in order to account for the principle of justice (Kubanyiova, 2015), it was ensured that the selected elementary schools, high schools, language schools, and the university were located in several different regions of Hungary, including both schools from the capital and schools from the countryside. This was ensured to avoid the selection of only privileged populations, namely, schools located in Budapest. The aim of the participant selection was to also include schools that might be less frequently researched and provide research benefit for them.

The research instruments and research strategies can also raise ethical dilemmas. In the present study, the main instruments used were the language proficiency tests. Because the digital platform was still in the development phase during the data collection, technical problems could still occur while the students were solving the tasks. These technical issues could include, for example, the sudden freezing of the software, data loss, or difficulties of recording an answer. Such problems could cause not only distress to the participant, but also possible permanent data loss in some cases. As the reliability and the validity of the test results is not only a methodological issue, but also the ethical right of the participant, in cases where some data loss was probably detected after the data collection, the participants' answers were disregarded completely during the data analyses. As customary, the opportunity was provided for the participants to opt out from the data collection at any time without any consequences, and thus those who were uncomfortable with the technical glitches and felt too distressed by them, could also leave.

Regarding the interviews and questionnaires used for the data collection, the main ethical concerns were also related to the previously mentioned principles. Regarding both

instruments, the participants were informed about the details of the data collection, their rights to withdraw, and the confidentiality, anonymity, and non-traceability of the data handling. These aspects were especially stressed in the case of the interviews and the think-aloud protocols because of the direct personal contact between the researcher and the participant. It was also emphasised that the participants had the right to refuse to answer any questions or withdraw completely at any point of the data collection without any consequence. This was especially important because several participants of the first and the second data collection phase were students of the researcher himself so they might have perceived the power relationship as an unequal one. Furthermore, during the data analysis, it was ensured that the interpretation of the data and the claims made based on them were reliable and valid, and the data provided by the participants was not intentionally misinterpreted. For the same reason, the interviews, the think-aloud protocols, and the questionnaire studies were conducted in Hungarian, the mother tongue of the participants, to allow for maximum freedom of expression for the participants and to avoid distortions of the data caused by the inappropriate skills of expression in a foreign language, especially in the case of the lower language proficiency levels.

5 Results and discussion

The aim of the following sections is to present and discuss the findings of the data collection. For the sake of a logical organisation, the results are presented and discussed in direct connection with the research questions. Section 5.1 (p. 75) discusses the data related to the first research question, section 5.2 (p. 107) contains the results related to the second research question, and section 5.3 (p. 118) presents and discusses the data related to the third research question.

5.1 Research question 1: Do the paper-based sets of tasks and the computer-based sets of tasks measure listening comprehension in an equally reliable way?

The aim of the first research question of the study was to investigate whether the sets of tasks administered in a paper-based format and those administered in a computer-based format measure in an equally reliable way. This question was raised because, as it has already been discussed in the theoretical background, the real-life practice of consuming multimedia material suggests that audio-only listening comprehension tasks do not necessarily represent the construct of listening comprehension to the full extent. For this reason, the computer-based test format was introduced in this research study to enable the use of audio-visual material in the test tasks.

As the real-life practice related to consuming multimedia material seem to encourage the supplementation of the listening comprehension component of language tests with the use of audio-visual material, it was essential to investigate whether the methodology supports such an endeavour. In the study, the participants were asked to first complete a language test in a paper-based format intended for their language proficiency level, then, at another occasion, they completed a similar test in a computer-based format. The two tests intended to measure the same language proficiency levels and they contained similar tasks (for a detailed description see Appendix 1A-16A). The only difference between them was

that in the computer-based test the last task was an audio-visual comprehension task as opposed to the audio-only tasks and the ATAO task in the paper-based test. In this way, if the two sets of tasks both seem to measure listening comprehension equally well, it can be proposed that the supplementation of the listening comprehension component of language tests with audio-visual tasks would not distort the language performance of candidates on the listening comprehension component. This supplementation would also be welcome as audio-only and audio-visual tasks together manage to reflect the real-life language use more authentically. In addition, the participants were also asked to fill in a questionnaire at the end of completing each set of tasks. The questionnaire intended to investigate the participants' perspectives about the potential distractors and difficulties experienced in connection with completing the tasks. The answers collected with these questionnaires were also taken into consideration in the present analysis of the reliability of the sets of tasks because it is could provide further insights into the possible issues.

5.1.1 Test results

The sets of tasks used in the present research study were designed as part of a larger language test development project, and the data analysed in this dissertation came from the pilot phase of that project. For this reason, these sets of tasks analysed in this section were not calibrated for the intended language proficiency levels yet. Therefore, the first step of the analysis was to investigate whether the items of the tests work appropriately. In order to do so, the Cronbach's alpha values of the sets of tasks were calculated along with the item facility values and the point-biserial correlations of each item in the sets of tasks with the help of SPSS 22.0 and Microsoft Excel to follow the classical test theory approach. The judgement of the Cronbach's alpha values was based on Kline (2000). For the Cronbach's alpha values, see Table 14.

Table 14

Cronbach's Alpha Values and Internal Consistency (Kline, 2000)

Cronbach's alpha	Internal consistency
$0.9 \leq \alpha$	Excellent
$0.8 \leq \alpha < 0.9$	Good
$0.7 \leq \alpha < 0.8$	Acceptable
$0.6 \leq \alpha < 0.7$	Questionable
$0.5 \leq \alpha < 0.6$	Poor
$\alpha < 0.5$	Unacceptable

Table 15

Reliability Measures of the English Paper-Based Tests

Proficiency level	Number of Items	Cronbach's Alpha	Mean	Standard Deviation	Variance	SEM
A2	24	0.71	16.91	3.27	10.69	1.67
B1	29	0.31	17.26	2.83	7.98	2.28
B2	29	0.52	18.62	3.47	12.01	2.35
C1	29	0.71	22.29	3.79	14.35	1.97

Note. Grey shading indicates problematic measures.

As Table 15 shows, the Cronbach's alpha values were acceptable for the A2 ($\alpha = 0.71$) and the C1 ($\alpha = 0.71$) level English paper-based tests; however, they were unacceptable for the B1 level test ($\alpha = 0.31$) and poor for the B2 level test ($\alpha = 0.52$). With the help of the SPSS 22.0 it was calculated that if items number 5, 12, 17, 19, 20, and 21 are deleted from the B2 test, the Cronbach's alpha value of the set becomes 0.74, which is considered to be acceptable. In the case of the B1 test, however, the calculations suggested that this set of tasks is poorly designed and the Cronbach's alpha value of the test cannot be improved.

In order to further investigate the important characteristics of the sets of tasks, the item facility values (i.e., item difficulty) and the point-biserial correlations for each item of the tests were calculated. The item facility value measures how easy or difficult an item is; whereas point-biserial correlations show how well an item can discriminate between the low and high performing students. Item facility values can range from 0.00 to 1.00. The ranges used for the item analysis in the present study were based on Brown and Hudson (2002) (see

Table 16). The *extremely difficult* items are marked as problematic for the investigated data set because such items are too challenging even for those test-takers who represent the top layer of the particular language proficiency band, and even these participants are usually unable to answer such items correctly. Similarly, the *very easy* items also threaten the reliability of the measurement because they can usually be answered correctly even by those test-takers who are at the bottom of or below the particular language proficiency level band. As such participants do not represent the intended target group of the test, the *extremely difficult* and *very easy* items should be disregarded. *Very difficult* items and moderately *easy items*, however, should be taken into consideration as they can elicit valuable information about the language knowledge of the test-taker. *Very difficult* items are challenging but not impossible to solve for the top test-takers, so they can distinguish the top test-takers from the average ones. On the other hand, *moderately easy* items are necessary in the test because they are relatively easy to solve for most test-takers; therefore, they can reduce the stress levels of the test-takers (Brown & Hudson, 2002).

Table 16

Item Facility Range (Brown & Hudson, 2002)

Range	**Label**
0.0-0.3	Extremely difficult
0.3-0.5	Very difficult
0.5-0.7	Moderately difficult
0.7-0.90	Moderately easy
0.90-1.0	Very easy

The point-biserial measure ranges from -1.0 to 1.0. The closer the value to 1.0, the better it can discriminate between low and high performing students. If the value is 0.00 the item cannot discriminate between the test-takers because it was either too difficult (i.e., no test-taker could answer it correctly) or too easy (i.e., all test-takers could answer it correctly). Negative discrimination measures are not appropriate either because they show that test-takers who scored high on the test overall answered the item incorrectly, and test-takers

who scored low on the test overall answered the item correctly. For large scale testing measures discrimination values between 0.09 and 0.30 are already in the acceptable and fairly good range; however, higher values are even better (Sheskin, 2011).

As Table 17 illustrates, there were several problematic items in the A2 level English paper-based test. The item facility values show that items 4, 7, 9, 10, 11, 16, 21, and 23 were very easy, so they could be solved even by those test-takers who did not necessarily possess an A2 level language proficiency yet. Therefore, these items should be reviewed. Similarly, items 13, 14, and 15 can be considered to be extremely difficult and impossible to solve even for those test-takers who belong to the top level of the A2 language proficiency band. These items should also be revised and modified in a later use of the test. Furthermore, the point-biserial correlations further highlight that items 4, 7, 9, 11, 13, 21, and 23 are problematic. In the case of items 4, 7, 9, 11, and 23 the point-biserial correlations are $r = 0.00$, which means that they are not discriminating between the low performing and high performing test-takers. This could be explained by the fact that these items proved to be very easy and they could be solved by all the test-takers. Item 13 is also a problematic item which does not discriminate between the low performing and high performing test-takers because it was extremely difficult, and it could not be solved by any test-takers. In addition, item 21 also fails to measure appropriately because the $r = -0.11$ value suggest that high performing test-takers achieved a low score on this item, whereas low performing students achieved a high score.

Table 17

A2 English Paper-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	0.36	0.25
	2	0.36	0.25
	3	0.82	0.71
	4	1.00	0.00
	5	0.73	0.48
	6	0.73	0.48
Task 2	**7**	1.00	0.00
	8	0.82	0.42
	9	1.00	0.00
	10	0.91	0.67
	11	1.00	0.00
	12	0.82	0.35
Task 3	**13**	0.00	0.00
	14	0.27	0.39
	15	0.18	0.30
	16	0.91	0.28
	17	0.82	0.42
	18	0.55	0.76
Task 4	**19**	0.46	0.14
	20	0.64	0.61
	21	0.91	-0.11
	22	0.82	0.35
	23	1.00	0.00
	24	0.82	0.42

Note. Grey shading indicates problematic measures.

In connection with the B1 English paper-based test, the item facility value and point-biserial calculations show that the problematic items were items 1, 2, 3, 9, 14, 17, 19, 21, 24, 25, 26, 27, 28 (see Table 18). Items 2, 21, 25, 27, and 28 proved to be very easy for those candidates who were supposed to be at the B1 language proficiency level. On the other hand, items 1, 3, and 14 seemed to be extremely difficult even for the top candidates. The point-biserial correlations suggest that items 19, 21, and 24 negatively discriminate between low performing and high performing test-takers, whereas item 25 does not discriminate at all because the item is too easy. Items 9, 17, and 26 could also be considered slightly problematic because their point-biserial correlation values are very close to 0. Based on their item facility values ($p = 0.58$, $p = 0.53$ and $p = 0.63$, respectively), these items are considered

to be moderately difficult items. For this reason, they might be excluded from further uses of the test, or they could be kept in their current form as the low point-biserial correlation values could be explained by the small sample size. With a larger sample, the point-biserial correlation values of items 9, 17, and 26 could be higher.

The Cronbach's alpha calculations for the B1 English paper-based test (see Table 15) also suggest that the items of this test do not measure appropriately. The Cronbach's alpha value for this set of items was $\alpha = 0.31$, and calculations suggest that, even by omitting the items which were flagged as problematic by the item facility value and point-biserial correlation calculations, the reliability of the test cannot be enhanced and the Cronbach's alpha value of the test cannot be further improved than $\alpha = 0.50$ which is considered to be poor reliability. Therefore, the B1 level English paper-based test should be substantially revised and its results are not considered in the further analysis in this study.

Table 18

B1 English Paper-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	0.11	0.15
	2	0.95	0.19
	3	0.16	0.32
	4	0.32	0.18
	5	0.63	0.15
	6	0.63	0.38
Task 2	**7**	0.68	0.26
	8	0.79	0.14
	9	0.58	0.04
	10	0.47	0.40
	11	0.37	0.47
	12	0.32	0.70
Task 3	**13**	0.53	0.24
	14	0.26	0.28
	15	0.31	0.62
	16	0.63	0.11
	17	0.53	0.01
	18	0.42	0.37
Task 4	**19**	0.84	-0.01
	20	0.79	0.09
	21	0.90	-0.27
	22	0.58	0.49
	23	0.47	0.25
	24	0.79	-0.23
Task 5	**25**	1.00	0.00
	26	0.63	0.07
	27	0.95	0.27
	28	0.90	0.15
	29	0.74	0.10

Note. Grey shading indicates problematic measures.

Considering the B2 English paper-based test, items 5, 7, 12, 16, 17, 19, 20, and 21 should be analysed more closely (see Table 19). Regarding the item facility values, item 7 appears to be very easy and item 20 appears to be extremely difficult. In addition, item 20 also seems to negatively discriminate between low performing and high performing candidates so it should be revised before using this set of tasks again. Similarly, items 5, 12, 17, 19, and 21 also show negative discrimination, so they should also be revised or completely omitted from the test. The point-biserial correlation of item 16 also suggests that the item is slightly problematic because the value $r = 0.07$ suggest low discrimination but,

as its item facility value indicates that it is a moderately difficult item, it can be kept without modification.

The Cronbach's alpha value of the B2 English paper-based test was originally $\alpha = 0.52$, but further calculations suggested that by omitting the items with negative discrimination (i.e., items 5, 12, 17, 19, 20, and 21) the Cronbach's alpha value of this set of tasks can be improved to $\alpha = 0.74$, which is considered to be acceptable reliability.

Table 19

B2 English Paper-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	0.62	0.32
	2	0.42	0.36
	3	0.62	0.16
	4	0.42	0.12
	5	0.46	-0.10
	6	0.77	0.47
Task 2	**7**	0.92	0.13
	8	0.89	0.34
	9	0.46	0.19
	10	0.62	0.25
	11	0.69	0.46
	12	0.89	-0.11
Task 3	**13**	0.81	0.73
	14	0.65	0.11
	15	0.42	0.18
	16	0.50	0.07
	17	0.81	-0.11
	18	0.85	0.66
Task 4	**19**	0.39	-0.12
	20	0.15	-0.48
	21	0.62	-0.18
	22	0.81	0.65
	23	0.73	0.61
	24	0.73	0.48
Task 5	**25**	0.73	0.48
	26	0.77	0.60
	27	0.62	0.41
	28	0.65	0.43
	29	0.62	0.44

Note. Grey shading indicates problematic measures.

Regarding the C1 level English paper-based test (see Table 20), the item facility values suggest that items 1, 2, 3, 5, 6, 9, 12, 18, and 21 are too easy and should be revised

and modified. Items 1, 3, 5, and 18 have $r = 0.00$ point-biserial correlations, which indicates that these items are not only too easy, but they also do not discriminate between low performing and high performing candidates at all. Furthermore, the point-biserial correlation values also show that items 14, 15, and 19 discriminate negatively among the candidates. Even though these items are moderately difficult, they should be deleted or revised because of the negative discrimination values. Furthermore, item 24 is also slightly problematic because it has a low discrimination value ($r = 0.06$); however, as its item facility value indicates that it is a moderately difficult item, it should be preserved in the test, nevertheless.

Table 20

C1 English Paper-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	1.00	0.00
	2	0.94	0.81
	3	1.00	0.00
	4	0.88	0.61
	5	1.00	0.00
	6	0.94	0.81
Task 2	**7**	0.77	0.15
	8	0.59	0.16
	9	0.94	0.81
	10	0.71	0.43
	11	0.82	0.69
	12	0.94	0.81
Task 3	**13**	0.82	0.52
	14	0.65	-0.17
	15	0.53	-0.02
	16	0.88	0.51
	17	0.65	0.51
	18	1.00	0.00
	19	0.59	-0.03
Task 4	**20**	0.88	0.56
	21	0.94	0.81
	22	0.65	0.45
	23	0.71	0.29
	24	0.65	0.06
Task 5	**25**	0.82	0.40
	26	0.41	0.25
	27	0.65	0.35
	28	0.41	0.44
	29	0.53	0.20

Note. Grey shading indicates problematic measures.

The Cronbach's alpha values shown in Table 21 indicate that the computer-based sets of items, namely, the A2 and B1 level tests have a good reliability ($\alpha = 0.90$ and $\alpha = 0.87$ respectively), the B2 set has an acceptable reliability ($\alpha = 0.78$) and the only problematic set is the C1 level test ($\alpha = 0.41$). The reliability of this set could only be improved by deleting eight items, which means that one-quarter of the test items does not measure appropriately and should be deleted. Because of the high number of erroneous items, the results provided by this set of tasks is not taken into consideration during the analysis in the present dissertation.

Table 21

Reliability Measures of the English Computer-Based Tests

Proficiency level	Number of Items	Cronbach's Alpha	Mean	Standard Deviation	Variance	SEM
A2	25	0.90	14.55	6.01	36.07	1.79
B1	28	0.87	19.95	5.35	28.61	1.87
B2	27	0.78	14.42	4.76	22.65	2.22
C1	32	0.41	23.29	3.16	9.97	2.35

Note. Grey shading indicates problematic measures.

As far as the A2 English computer-based test is concerned (see Table 22), the item facility values suggest that items 8, 15, and 24 are very easy, while items 3, 5, 6, and 7 are extremely difficult. Regarding the point-biserial correlation values item 5 (p = 0.00) does not discriminate among low and high performing test-takers, and the discrimination power of item 3 (p = 0.03) is also insufficient. The point-biserial correlation of item 20 (p = -0.12) discriminate negatively between low and high performing students.

Table 22

A2 English Computer-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	0.73	0.70
	2	0.46	0.49
	3	0.09	0.08
	4	0.82	0.67
	5	0.00	0.00
	6	0.27	0.52
Task 2	**7**	0.18	0.47
	8	0.91	0.77
	9	0.82	0.83
	10	0.55	0.63
	11	0.46	0.31
	12	0.64	0.51
	13	0.46	0.34
Task 3	**14**	0.64	0.70
	15	0.91	0.77
	16	0.73	0.43
	17	0.36	0.37
	18	0.82	0.83
	19	0.55	0.48
Task 4	**20**	0.27	-0.12
	21	0.73	0.43
	22	0.82	0.83
	23	0.64	0.64
	24	0.91	0.77
	25	0.82	0.83

Note. Grey shading indicates problematic measures.

The B1 level English computer-based test contained eleven problematic items (see Table 23). Items 2, 3, 6, 7, 9, 11, 24, 25, and 28 were very easy, and items 1 (p = 0.21) and 23 (p = 0.21) were extremely difficult. Nevertheless, based on the point-biserial correlations even these items are discriminating appropriately between the low performing and high performing candidates. Therefore, they do not have to be disregarded in the analysis.

Table 23

B1 English Computer-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	0.21	0.05
	2	0.90	0.61
	3	0.95	0.88
	4	0.79	0.60
	5	0.79	0.48
	6	0.95	0.88
Task 2	**7**	0.95	0.88
	8	0.63	0.48
	9	0.90	0.73
	10	0.42	0.27
	11	0.90	0.70
	12	0.84	0.62
Task 3	**13**	0.53	0.29
	14	0.79	0.60
	15	0.69	0.37
	16	0.74	0.53
	17	0.37	0.21
Task 4	**18**	0.63	0.44
	19	0.84	0.64
	20	0.74	0.53
	21	0.32	0.11
	22	0.63	0.40
	23	0.21	0.13
Task 5	**24**	0.95	0.88
	25	0.95	0.88
	26	0.63	0.26
	27	0.84	0.64
	28	0.90	0.57

Note. Grey shading indicates problematic measures.

Considering the B2 English computer-based test (see Table 24), only items 15 (p = 0.19) and 18 (p = 0.19) were very easy, on the basis of their item facility values. However, they do discriminate in an appropriate way. In addition to this, only item 23 had insufficient discrimination power (r = 0.01); thus, this item should be deleted or redesigned.

Table 24

B2 English Computer-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	0.65	0.42
	2	0.62	0.40
	3	0.73	0.33
	4	0.50	0.54
	5	0.65	0.25
	6	0.50	0.53
Task 2	**7**	0.27	0.29
	8	0.73	0.45
	9	0.81	0.37
	10	0.77	0.14
	11	0.81	0.58
Task 3	**12**	0.81	0.51
	13	0.81	0.47
	14	0.54	0.34
	15	0.19	0.53
	16	0.23	0.60
	17	0.31	0.41
Task 4	**18**	0.19	0.43
	19	0.35	0.36
	20	0.31	0.43
	21	0.46	0.32
	22	0.27	0.13
Task 5	**23**	0.50	0.01
	24	0.77	0.20
	25	0.39	0.48
	26	0.58	0.34
	27	0.69	0.34

Note. Grey shading indicates problematic measures.

Concerning the C1 English computer-based test, the item facility value and point-biserial calculations indicate that the majority of the items are problematic (see Table 25). Items 5, 19, and 31 appeared to be very easy for the test-takers. Moreover, the point-biserial correlations of items 5 and 31 also show negative discrimination, so these items should be deleted from the test. Similarly, the point-biserial correlations of items 17, 18, 22, 24, 26, and 32 also discriminate negatively between the low performing and high performing test-takers. Despite the fact that these items are moderately difficult to moderately easy, because of their negative discrimination power, they should be omitted. Items 2 and 15 also have a low discrimination power (r = 0.03 and r = 0.01 respectively).

The Cronbach's alpha value for the C1 English computer-based test (see Table 21) appears to also highlight the fact that this set of tasks does not measure C1 level language proficiency appropriately. The Cronbach's alpha value of the test was $\alpha = 0.41$ and only by deleting eight items could the acceptable reliability of $\alpha = 0.71$ be achieved. This would mean that one-quarter of the items must be deleted in order for the set of tasks to measure appropriately. Therefore, the C1 English computer-based test should be thoroughly revised and redesigned.

Table 25

C1 English Computer-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	0.82	0.39
	2	0.88	0.03
	3	0.82	0.48
	4	0.71	0.43
	5	0.94	-0.06
	6	0.82	0.73
Task 2	**7**	0.77	0.62
	8	0.59	0.31
	9	0.82	0.58
	10	0.35	0.59
	11	0.65	0.54
	12	0.47	0.40
Task 3	**13**	0.77	0.49
	14	0.59	0.34
	15	0.77	0.01
	16	0.59	0.31
	17	0.65	-0.01
	18	0.88	-0.02
Task 4	**19**	0.94	0.18
	20	0.77	0.23
	21	0.71	0.35
	22	0.88	-0.14
	23	0.82	0.14
	24	0.53	-0.29
	25	0.82	0.24
	26	0.77	-0.21
Task 5	**27**	0.59	0.38
	28	0.71	0.14
	29	0.59	0.19
	30	0.82	0.14
	31	0.94	-0.21
	32	0.53	-0.47

Note. Grey shading indicates problematic measures.

As Table 26 shows, the Cronbach's alpha values were not acceptable for any of the proficiency levels as far as the German paper-based tests are concerned. However, with the help of the SPSS 22.0 it was calculated that if items number 1, 4, 14, 16, and 18 are deleted from the A2 test, the Cronbach's alpha value of the set becomes 0.72; if items 18, 19, 24, and 25 are omitted from the B1 test, the Cronbach's alpha value of the set becomes 0.71; and if items 2 and 8 are deleted from the C1 test, the Cronbach's alpha value of the set

becomes 0.71. In the case of the B2 test, however, the calculations suggest that the set of tasks is poorly designed and the Cronbach's alpha value of the test cannot be improved.

Table 26

Reliability Measures of the German Paper-Based Tests

Proficiency level	Number of Items	Cronbach's Alpha	Mean	Standard Deviation	Variance	SEM
A2	24	0.54	17.82	2.68	7.16	1.74
B1	29	0.59	14.63	3.77	14.25	2.37
B2	29	0.48	15.04	3.58	12.82	2.52
C1	29	0.60	19.23	3.54	12.53	2.16

Note. Grey shading indicates problematic measures.

As far as the item facility values are concerned, in the A2 German paper-based test (see Table 27), items 2, 3, 5, 6, 11, 12, 15, 16, 17, 20, 21, and 23 were very easy, however for items 2, 3, 6, and 23 the point-biserial correlations are acceptable. Nevertheless, items 5, 11, 12, 15, 17, 20, and 21 do not discriminate at all between low and high performing test-takers. In addition to this, item 16 has a negative discrimination value. Item 13 appears to be an extremely difficult item ($p = 0.27$) with a good discriminating power, while item 1 has a low discrimination power ($r = 0.02$) with a moderately difficult item facility index.

Table 27

A2 German Paper-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	0.67	0.02
	2	0.91	0.57
	3	0.91	0.57
	4	0.55	0.14
	5	1.00	0.00
	6	0.91	0.21
Task 2	**7**	0.64	0.23
	8	0.36	0.26
	9	0.64	0.51
	10	0.55	0.35
	11	1.00	0.00
	12	1.00	0.00
Task 3	**13**	0.27	0.58
	14	0.64	0.23
	15	1.00	0.00
	16	0.91	-0.26
	17	1.00	0.00
	18	0.46	0.27
Task 4	**19**	0.55	0.35
	20	1.00	0.00
	21	1.00	0.00
	22	0.27	0.58
	23	0.91	0.57
	24	0.73	0.64

Note. Grey shading indicates problematic measures.

Concerning the item facility values of the B1 German paper-based test (see Table 28), items 3, 7, 12, and 14 appear to be extremely difficult with good discrimination power. As far as the discrimination power is concerned, however, items 10, 17, 19, and 25 have low discrimination values, and items 18 and 24 discriminate negatively (r = -0.04 and r = -0.18, respectively). Therefore, the latter two items should be completely disregarded from the test.

Table 28

B1 German Paper-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	0.63	0.36
	2	0.58	0.59
	3	0.21	0.60
	4	0.84	0.23
	5	0.74	0.10
	6	0.31	0.25
Task 2	**7**	0.26	0.66
	8	0.32	0.37
	9	0.74	0.45
	10	0.79	0.05
	11	0.53	0.47
	12	0.11	0.31
Task 3	**13**	0.95	0.23
	14	0.16	0.62
	15	0.32	0.43
	16	0.63	0.10
	17	0.63	0.01
	18	0.68	-0.04
Task 4	**19**	0.79	0.02
	20	0.58	0.11
	21	0.47	0.40
	22	0.37	0.42
	23	0.32	0.52
	24	0.53	-0.18
Task 5	**25**	0.26	0.03
	26	0.32	0.49
	27	0.63	0.01
	28	0.53	0.30
	29	0.42	0.20

Note. Grey shading indicates problematic measures.

The item facility values in the B2 German paper-based test show that items 3, 7, 9, and 20 appear to be extremely difficult (see Table 29). However, while items 7, 9, and 20 have acceptable discrimination power, item 3 has a weak discrimination value (r = 0.08). Similarly, items 2 and 22 have low discrimination values with a moderately difficult and a moderately easy item facility value respectively. Items 1, 8, and 11, however, discriminate negatively (r = -0.01, r = -0.01, and r = -0.08, respectively); therefore, they should be completely disregarded from the test.

Table 29

B2 German Paper-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	0.38	-0.01
	2	0.58	0.06
	3	0.21	0.08
	4	0.42	0.25
	5	0.58	0.29
	6	0.63	0.13
Task 2	**7**	0.25	0.37
	8	0.38	-0.01
	9	0.25	0.18
	10	0.58	0.20
	11	0.38	-0.08
	12	0.54	0.22
Task 3	**13**	0.71	0.21
	14	0.54	0.17
	15	0.46	0.50
	16	0.50	0.50
	17	0.58	0.27
	18	0.54	0.62
Task 4	**19**	0.46	0.41
	20	0.29	0.33
	21	0.67	0.38
	22	0.75	0.03
	23	0.79	0.18
	24	0.50	0.15
Task 5	**25**	0.50	0.13
	26	0.54	0.50
	27	0.58	0.34
	28	0.79	0.49
	29	0.67	0.30

Note. Grey shading indicates problematic measures.

The item facility values in the C1 German paper-based test demonstrate that items 3, 11, 12, 17, 19, and 27 were very easy; however, items 11 and 27 had an acceptable level of point-biserial correlation (r = 0.75 and r = 0.10, respectively) (see Table 30). On the other hand, items 3, 12, 17 had no discrimination power, and item 19 had a low discrimination value. Regarding some other discrimination values, items 2, 4, 7, 8, 25 and 26 had negative discrimination power; thus, these items need to be redesigned before further use.

Table 30

C1 German Paper-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	0.77	0.40
	2	0.46	-0.10
	3	1.00	0.00
	4	0.39	-0.01
	5	0.69	0.33
	6	0.46	0.51
Task 2	**7**	0.92	-0.14
	8	0.15	-0.75
	9	0.77	0.55
	10	0.69	0.09
	11	0.92	0.75
	12	1.00	0.00
Task 3	**13**	0.46	0.29
	14	0.77	0.35
	15	0.69	0.66
	16	0.46	0.51
	17	1.00	0.00
	18	0.77	0.60
	19	0.92	0.02
Task 4	**20**	0.77	0.50
	21	0.54	0.63
	22	0.31	0.52
	23	0.46	0.42
	24	0.85	0.75
Task 5	**25**	0.62	-0.13
	26	0.46	-0.10
	27	0.92	0.10
	28	0.31	0.38
	29	0.69	0.33

Note. Grey shading indicates problematic measures.

The Cronbach's alpha values of the German computer-based tests suggests that the B1 and C1 level tests are measuring appropriately, their Cronbach's alpha values being $\alpha = 0.71$ and $\alpha = 0.79$, which both count as acceptable levels of reliability (see Table 31). In contrast, the Cronbach's alpha value of the A2 level test is $\alpha = 0.52$, which is poor reliability. However, by eliminating items 2, 13, and 21, the Cronbach's alpha value can be increased to $\alpha = 0.71$. Similarly, the original Cronbach's alpha value of the B2 level test indicates only a questionable level of reliability ($\alpha = 0.63$), but by excluding items 8, 21 and 27, the reliability of the test becomes acceptable ($\alpha = 0.70$). These results suggest that all the sets of

tasks designed for the German computer-based tests managed to measure the intended language proficiency levels adequately, and only a few items have to be omitted from the analysis.

Table 31

Reliability Measures of the German Computer-Based Tests

Proficiency level	Number of Items	Cronbach's Alpha	Mean	Standard Deviation	Variance	SEM
A2	25	0.52	21.82	2.23	4.96	1.47
B1	28	0.71	15.84	4.22	17.81	2.20
B2	27	0.63	14.83	3.88	15.01	2.32
C1	32	0.79	23.85	4.91	24.14	2.15

Note. Grey shading indicates problematic measures.

When analysing the A2 German computer-based test (Table 32), the item facility values suggest that 14 out of the 25 items were very easy for the candidates, and 11 out of these 14 very easy items do not have any discrimination power, their point-biserial correlations being $r = 0.00$. These items should be thoroughly redesigned, as in their current state, they do not provide appropriate information about the language competence of the candidates. In contrast, the item facility values of items 11, 24, and 25 also indicate that these items are very easy; however, their point-biserial correlation values suggest that they still have a discrimination power, so they might be preserved in their current state. Other problematic items which should be revised or deleted from the test are items 2 and 21. These items have a negative discrimination power, and they can skew the results of the test. The need for the deletion of these items is further reinforced by the Cronbach's alpha value of the scale as, in order to achieve an acceptable level of reliability, the said items should be omitted.

Table 32

A2 German Computer-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	1.00	0.00
	2	0.55	-0.07
	3	0.82	0.38
	4	1.00	0.00
	5	0.82	0.49
	6	0.64	0.70
Task 2	**7**	1.00	0.00
	8	1.00	0.00
	9	1.00	0.00
	10	1.00	0.00
	11	0.91	0.40
	12	1.00	0.00
	13	0.73	0.41
Task 3	**14**	0.82	0.17
	15	0.55	0.34
	16	0.73	0.32
	17	1.00	0.00
	18	1.00	0.00
	19	0.82	0.49
Task 4	**20**	1.00	0.00
	21	0.82	-0.14
	22	0.82	0.49
	23	1.00	0.00
	24	0.91	0.68
	25	0.91	0.68

Note. Grey shading indicates problematic measures.

The analysis of the B1 German computer-based test indicates the presence of fewer problematic items than the A2 German computer-based test (Table 33). Four items, namely items 1, 5, 10, and 24, proved to be very easy. Items 1 and 10 also have no discrimination power, their point-biserial correlation values being $r = 0.00$; whereas, items 5 and 24 have a negative discrimination power, their point-biserial values being $r = -0.06$. Items 13, 16, 19, and 28 were extremely difficult. Item 28 has a negative discrimination power ($r = -0.12$), and item 19 has a low discrimination power, so these items should be deleted from the test. Items 13 and 16 could be kept without modifications because they have some discrimination power. Other items which should be revised or modified because of their discrimination power are items 18, 23 and 27.

Table 33

B1 German Computer-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	1.00	0.00
	2	0.63	0.59
	3	0.47	0.41
	4	0.68	0.27
	5	0.95	-0.06
	6	0.58	0.73
Task 2	**7**	0.32	0.56
	8	0.74	0.12
	9	0.84	0.43
	10	1.00	0.00
	11	0.58	0.47
	12	0.58	0.40
Task 3	**13**	0.21	0.39
	14	0.37	0.52
	15	0.47	0.59
	16	0.05	0.51
	17	0.63	0.54
Task 4	**18**	0.74	-0.05
	19	0.26	0.05
	20	0.42	0.41
	21	0.53	0.54
	22	0.53	0.56
	23	0.63	0.07
Task 5	**24**	0.95	-0.06
	25	0.37	0.26
	26	0.53	0.41
	27	0.53	0.01
	28	0.26	-0.12

Note. Grey shading indicates problematic measures.

The B2 German computer-based test (Table 34) was the other German computer-based set of tasks besides the A2 one where the initial Cronbach's alpha values of the test suggested questionable reliability. This fact is also supported by the number of problematic items in the test. Item 7 appears to be too easy with an item facility value of $p = 0.96$, with a low discrimination power ($r = 0.04$). Items 5, 11, 14 and 23 are extremely difficult items; however, they discriminate between low and high performing test-takers so they do not necessarily have to be modified in later uses of the test. In contrast, items 8, 21 and 27 have negative discrimination power and they should be deleted. The Cronbach's

alpha calculations also support the elimination of these items from the test because, by their omission, the reliability of the test can be increased from $\alpha = 0.63$ to $\alpha = 0.70$.

Table 34

B2 German Computer-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	0.58	0.51
	2	0.46	0.19
	3	0.71	0.11
	4	0.75	0.37
	5	0.29	0.64
	6	0.68	0.36
Task 2	**7**	0.96	0.04
	8	0.83	-0.08
	9	0.50	0.39
	10	0.63	0.32
	11	0.25	0.37
Task 3	**12**	0.46	0.64
	13	0.33	0.30
	14	0.13	0.63
	15	0.54	0.35
	16	0.63	0.14
	17	0.71	0.07
Task 4	**18**	0.46	0.38
	19	0.71	0.61
	20	0.50	0.62
	21	0.50	-0.04
	22	0.79	0.24
Task 5	**23**	0.25	0.22
	24	0.50	0.22
	25	0.50	0.30
	26	0.63	0.10
	27	0.58	-0.01

Note. Grey shading indicates problematic measures.

In the case of the C1 German computer-based test (Table 35), there were several items which are considered to be too easy, on the basis of their item facility values. These items are items 3, 7, 11, 13, 16 and 31. Out of these items, items 11 and 13 should be substantially revised because the also have no discrimination values. In contrast, items 3 and 31 can be preserved in the test without revision because they still have discrimination power ($r = 0.58$ and $r = 0.11$ respectively) although their item facility values indicate that they are

too easy (p = 0.92). Items 7 and 16 appear to be very problematic as they are not only too easy, but they also have a negative discrimination power. Furthermore, items 2, 8, and 24 also have a negative discrimination power so they must be redesigned. Item 18 might also be redesigned because it has a rather low discrimination power (r = 0.06) despite the fact that it is a very difficult item (p = 0.39).

Table 35

C1 German Computer-Based Test: Item Facility Values and Point-Biserial Correlations

Task No.	Item No.	IF	r_{pbi}
Task 1	**1**	0.85	0.46
	2	0.69	-0.16
	3	0.92	0.58
	4	0.85	0.85
	5	0.85	0.20
	6	0.54	0.54
Task 2	**7**	0.92	-0.07
	8	0.85	-0.14
	9	0.69	0.25
	10	0.85	0.29
	11	1.00	0.00
	12	0.77	0.17
Task 3	**13**	1.00	0.00
	14	0.31	0.26
	15	0.69	0.49
	16	0.92	-0.07
	17	0.85	0.85
	18	0.39	0.06
Task 4	**19**	0.77	0.76
	20	0.77	0.80
	21	0.54	0.41
	22	0.54	0.44
	23	0.62	0.10
	24	0.77	-0.20
	25	0.77	0.80
	26	0.85	0.51
Task 5	**27**	0.77	0.50
	28	0.69	0.69
	29	0.85	0.85
	30	0.54	0.22
	31	0.92	0.11
	32	0.54	0.41

Note. Grey shading indicates problematic measures.

To conclude, the results of the Cronbach's alpha analysis, the item facility values, and the point-biserial correlations of the English and German paper-based and computer-based tests suggest that the majority of the tests manage to measure the intended language proficiency levels in a satisfactory way for the present research purposes. In spite of having negative point-biserial values in many of the tests, as the Cronbach's alpha values of these sets of tasks suggest a satisfactory level of reliability, the results of these tests are

taken into consideration during answering the proposed research questions. Based on their Cronbach's alpha values, the A2 and C1 English paper-based tests had a satisfactory reliability value without modifications, whereas the B2 English paper-based test was not satisfactory but its reliability value could be improved by deleting certain items. However, the B1 English paper-based test does not measure appropriately and has to be excluded from the rest of the analysis. Similarly, in the case of the English computer-based test, the A2, B1, and B2 levels managed to measure the intended language proficiency levels in a satisfactory way; whereas the C1 level test is unreliable and has to be omitted in the rest of the analysis. In connection with the German paper-based tests, none of the sets of tasks provided reliable measurement initially. Nevertheless, by deleting certain items, the A2, B1, and C1 tests could be improved to a satisfactory level of reliability. The B2 test, however, has to be excluded from the rest of the analysis. The B1 and C1 German computer-based tests showed a satisfactory level of reliability, whereas the A2 and B2 German computer-based tests could be improved by deleting some items.

5.1.2 Questionnaire results

The questionnaire was administered to the participants at the end of each test, and there were two different questionnaire versions designed for the paper-based and the computer-based tests. At both test versions, the participants received the questionnaire in a printed format, and they had approximately 15 minutes to fill it in. The questionnaire intended to investigate four constructs: (1) possible *disturbing features of the test* related to such acoustic and prosodic elements as the accent of the speakers or the background music for example; (2) issues related to the *structure of the test*, such as the number of items and number of listening opportunities; (3) the *perceived difficulty of the test*, for example, the existence of unknown vocabulary; and (4) the *necessity or relevance of having a video* in the last task. For the full questionnaires created for the paper-based and the computer-based

test, see Appendix 1C-2D. The present analysis only focuses on the results of the first three constructs as the last construct is discussed in connection with the third research question.

In order to analyse the participants' answers, the mean values and the standard deviations of each construct were calculated with SPSS 22.0 for each test version and proficiency level (see Table 36). Regarding the presence of *disturbing features* in the test, the majority of the participants at every language proficiency level seem to agree that they did not perceive any major disturbing factors in the test. The highest mean values for this construct can be found related to the A2 English paper-based test (M = 2.82, SD = 1.28) and the B1 German paper-based test (M = 2.82, SD = 0.75); however, even these values suggest that the majority of the participants did not find the acoustic and prosodic features of the test very disturbing. The majority of the participants at every proficiency level also seem to agree that the structure of the tests is satisfactory. The lowest mean value occurs at the B2 German paper-based test (M = 3.34, SD = 0.50), which still indicates moderate satisfaction with the test structure. In connection with the *perceived difficulty of the test*, the results are more varied across the different language proficiency levels. The highest level of difficulty was indicated at the B2 German paper-based test (M = 4.12, SD = 0.61) which suggests that the majority of the participants found the test tasks difficult. The least difficult set of tasks appeared to be the A2 German computer-based test (M = 2.31, SD = 0.54). For the rest of the sets of tasks, based on the mean values, the participants indicated a moderate to low difficulty of the test.

Table 36

Test Questionnaires: Descriptive Statistics

Proficiency level	Language	Test version	Construct	Mean	Std. Deviation
A2	English	Paper-based	Disturbing features of the test	2.82	1.28
			Structure of the test	3.53	0.73
			Perceived difficulty of the test	3.60	0.80
		Computer-based	Disturbing features of the test	1.93	0.50
			Structure of the test	3.49	0.96
			Perceived difficulty of the test	3.11	0.92
	German	Paper-based	Disturbing features of the test	2.55	1.03
			Structure of the test	4.35	0.46
			Perceived difficulty of the test	2.95	0.83
		Computer-based	Disturbing features of the test	1.48	0.48
			Structure of the test	4.58	0.53
			Perceived difficulty of the test	2.31	0.54
B1	English	Paper-based	Disturbing features of the test	2.82	0.90
			Structure of the test	4.20	0.51
			Perceived difficulty of the test	2.99	0.94
		Computer-based	Disturbing features of the test	1.95	0.97
			Structure of the test	4.40	0.74
			Perceived difficulty of the test	2.68	0.76
	German	Paper-based	Disturbing features of the test	2.82	0.75
			Structure of the test	3.59	0.60
			Perceived difficulty of the test	3.61	0.60
		Computer-based	Disturbing features of the test	2.14	0.79
			Structure of the test	3.87	0.65
			Perceived difficulty of the test	3.34	0.75

Proficiency level	Language	Test version	Construct	Mean	Std. Deviation
B2	English	Paper-based	Disturbing features of the test	2.36	0.87
			Structure of the test	4.23	0.71
			Perceived difficulty of the test	2.75	0.97
		Computer-based	Disturbing features of the test	2.27	0.87
			Structure of the test	3.82	0.79
			Perceived difficulty of the test	2.65	0.81
	German	Paper-based	Disturbing features of the test	2.57	0.76
			Structure of the test	3.34	0.50
			Perceived difficulty of the test	4.12	0.61
		Computer-based	Disturbing features of the test	1.89	0.74
			Structure of the test	3.97	0.57
			Perceived difficulty of the test	3.21	0.67
C1	English	Paper-based	Disturbing features of the test	2.26	0.68
			Structure of the test	4.34	0.52
			Perceived difficulty of the test	2.32	0.55
		Computer-based	Disturbing features of the test	2.26	0.34
			Structure of the test	3.93	0.62
			Perceived difficulty of the test	2.64	0.74
	German	Paper-based	Disturbing features of the test	2.00	0.69
			Structure of the test	4.15	0.49
			Perceived difficulty of the test	3.03	0.80
		Computer-based	Disturbing features of the test	1.62	0.71
			Structure of the test	4.54	0.31
			Perceived difficulty of the test	2.65	0.66

5.1.3 Conclusion

Considering the test results, it can be concluded that the majority of the designed sets of tasks manage to measure the intended language proficiency levels in a satisfactory way. The item facility values and the point-biserial correlations of the test items suggest that all tests except for the B1 English paper-based, the C1 English computer-based, and the B2 German paper-based tests manage to measure language proficiency in a reliable way. Combining the results of the questionnaires and the results of the tests, it can be claimed that there is a possibility of supplementing the listening comprehension component of language proficiency tests with audio-visual tasks because the reliability of the tests and the participants' perceptions about the different features of the test imply that there was no major difference between the results of the paper-based and the results of the computer-based tests. Even though there are shortcoming is terms of the reliability of some of the sets of tasks, it can still be claimed that the overall results are not worse on the computer-based than on the paper-based tests; therefore, the computer-based sets of tasks do not measure participants' listening skills in a less reliable way than the paper-based sets of tasks. As, based on the discussion so far, the real-world context appears to necessitate it, and the methodology does not seem to go against it, the supplementation of the listening comprehension component of language tests with an audio-visual task should be taken into consideration. In order to further support this claim, the following section analyses and discusses the differences between the participants' results achieved on the final tasks of the paper-based and the computer-based tests.

5.2 Research question 2: Does the performance of the test-takers on the audio-visual-to-audio-only tasks differ from their performance on the audio-visual tasks?

In both the paper-based and the computer-based tests, the last task was a task originally created from audio-visual material. However, in the case of the paper-based tests,

the recording of the last task had to be modified by removing the visual material from the originally audio-visual recording. The present dissertation uses the term *audio-visual-to-audio-only (ATAO)* task to refer to such tasks. The removal of the visual material was necessary because of feasibility reasons as during the administration of the paper-based tests only CD-players were available for playing the recordings. In the computer-based tests, however, the test-takers had the opportunity to watch the audio-visual material on laptops so the visual material of the last task could be played. The comparison of the participants' results on these two different tasks could shed light on the extent to which their performance might be influenced by the presence of the visual material. For the detailed results, see Table 37 - Table 44.

Table 37

Comparison of the Participants' Results on the Last Tasks in the English A2 Paper-Based and Computer-Based Tests

		Item No.										Item No.							
Test version	**Candidate No.**	**19**	**20**	**21**	**22**	**23**	**24**	**Σ**	**%**	**Test version**	**Candidate No.**	**20**	**21**	**22**	**23**	**24**	**25**	**Σ**	**%**
paper-based	1a	0	1	1	1	1	1	5	83	**computer-based**	1b	0	1	1	1	1	1	5	83
	2a	1	0	1	1	1	1	5	83		2b	0	1	1	1	1	1	5	83
	3a	0	1	1	1	1	1	5	83		3b	0	1	1	0	1	1	4	67
	4a	1	1	1	1	1	1	6	100		4b	0	0	1	1	1	1	4	67
	5a	1	1	1	0	1	1	5	83		5b	0	1	1	0	1	1	4	67
	6a	0	1	1	1	1	1	5	83		6b	0	1	1	1	1	1	5	83
	7a	1	1	1	1	1	1	6	100		7b	1	0	1	1	1	1	5	83
	8a	0	0	1	1	1	1	4	67		8b	0	0	0	0	0	0	0	0
	9a	1	0	1	0	1	0	3	50		9b	1	1	1	1	1	1	6	100
	10a	0	0	1	1	1	1	4	67		10b	1	1	0	0	1	0	3	50
	11a	0	1	0	1	1	0	3	50		11b	0	1	1	1	1	1	5	83
		Average %							**77**			**Average %**							**70**

Note. Light grey shading indicates higher performance. Dark grey shading indicates average percentages.

Table 38

Comparison of the Participants' Results on the Last Tasks in the English B1 Paper-Based and Computer-Based Tests

		Item No.									Item No.						
Test version	**Candidate No.**	**25**	**26**	**27**	**28**	**29**	**Σ**	**%**	**Test version**	**Candidate No.**	**24**	**25**	**26**	**27**	**28**	**Σ**	**%**
paper-based	12a	1	1	1	1	0	4	80	**computer-based**	12b	1	1	0	1	1	4	80
	13a	1	0	0	1	0	2	40		13b	1	1	0	1	1	4	80
	14a	1	0	1	1	1	4	80		14b	1	1	0	1	1	4	80
	15a	1	0	1	1	1	4	80		15b	1	1	1	1	1	5	100
	16a	1	0	1	0	1	3	60		16b	1	1	1	1	1	5	100
	17a	1	1	1	1	1	5	100		17b	1	1	1	1	1	5	100
	18a	1	1	1	1	1	5	100		18b	1	1	1	1	1	5	100
	19a	1	1	1	1	1	5	100		19b	1	1	1	1	1	5	100
	20a	1	1	1	1	1	5	100		20b	1	1	0	1	0	3	60
	21a	1	1	1	1	0	4	80		21b	0	0	0	0	0	0	0
	22a	1	1	1	1	1	5	100		22b	1	1	0	1	1	4	80
	23a	1	1	1	1	1	5	100		23b	1	1	1	1	1	5	100
	24a	1	1	1	1	0	4	80		24b	1	1	1	1	1	5	100
	25a	1	1	1	0	1	4	80		25b	1	1	1	1	1	5	100
	26a	1	1	1	1	0	4	80		26b	1	1	1	1	1	5	100
	27a	1	0	1	1	1	4	80		27b	1	1	1	1	1	5	100
	28a	1	0	1	1	1	4	80		28b	1	1	0	1	1	4	80
	29a	1	0	1	1	1	4	80		29b	1	1	1	0	1	4	80
	30a	1	1	1	1	1	5	100		30b	1	1	1	0	1	4	80
		Average %						**84**			**Average %**						**85**

Note. Light grey shading indicates higher performance. Dark grey shading indicates average percentages. The paper-based version does not have an acceptable Cronbach's alpha value.

Table 39

Comparison of the Participants' Results on the Last Tasks in the English B2 Paper-Based and Computer-Based Tests

		Item No.									Item No.						
Test version	**Candidate No.**	**25**	**26**	**27**	**28**	**29**	**Σ**	**%**	**Test version**	**Candidate No.**	**23**	**24**	**25**	**26**	**27**	**Σ**	**%**
paper-based	31a	1	1	1	1	1	5	100	**computer-based**	31b	0	0	1	1	1	3	60
	32a	1	1	0	0	1	3	60		32b	0	1	1	1	1	4	80
	33a	1	1	1	1	1	5	100		33b	0	1	1	1	1	4	80
	34a	1	1	1	1	1	5	100		34b	1	1	0	1	1	4	80
	35a	1	1	1	1	0	4	80		35b	1	1	1	1	0	4	80
	36a	1	1	1	0	1	4	80		36b	1	0	0	1	0	2	40
	37a	1	1	1	1	1	5	100		37b	0	0	0	0	1	1	20
	38a	1	1	1	1	0	4	80		38b	1	1	1	1	1	5	100
	39a	0	1	1	0	1	3	60		39b	1	1	0	0	1	3	60
	40a	1	1	0	1	1	4	80		40b	1	1	1	1	0	4	80
	41a	1	1	1	1	1	5	100		41b	0	1	0	1	0	2	40
	42a	1	1	1	1	0	4	80		42b	0	1	0	0	1	2	40
	43a	1	1	1	1	0	4	80		43b	0	0	0	1	1	2	40
	44a	1	1	0	1	0	3	60		44b	1	1	0	0	1	3	60
	45a	1	1	1	1	1	5	100		45b	0	0	0	0	1	1	20
	46a	0	0	0	0	0	0	0		46b	0	1	0	1	1	3	60
	47a	0	1	0	0	1	2	40		47b	1	1	0	1	1	4	80
	48a	1	0	0	1	0	2	40		48b	1	1	0	1	0	3	60
	49a	1	1	1	1	1	5	100		49b	0	0	0	0	0	0	0
	50a	1	1	1	1	0	4	80		50b	1	1	0	0	1	3	60
	51a	0	0	1	0	1	2	40		51b	0	1	1	0	1	3	60
	52a	0	0	0	1	1	2	40		52b	0	1	1	0	0	2	40
	53a	0	0	0	0	0	0	0		53b	0	1	0	0	1	2	40
	54a	0	0	0	0	0	0	0		54b	1	1	1	0	0	3	60
	55a	1	1	1	1	1	5	100		55b	1	1	0	1	1	4	80
	56a	1	1	0	0	1	3	60		56b	1	1	1	1	1	5	100
		Average %						**68**			**Average %**						**58**

Note. Light grey shading indicates higher performance. Dark grey shading indicates average percentages.

Table 40

Comparison of the Participants' Results on the Last Tasks in the English C1 Paper-Based and Computer-Based Tests

		Item No.									Item No.							
Test version	**Candidate No.**	**25**	**26**	**27**	**28**	**29**	**Σ**	**%**	**Test version**	**Candidate No.**	**27**	**28**	**29**	**30**	**31**	**32**	**Σ**	**%**
paper-based	57a	1	0	0	0	1	2	40	**computer-based**	57b	1	1	1	1	0	0	4	67
	58a	1	1	1	1	1	5	100		58b	1	1	1	1	1	0	5	83
	59a	1	0	1	0	0	2	40		59b	1	1	1	1	1	0	5	83
	60a	1	0	0	1	0	2	40		60b	1	1	1	1	1	0	5	83
	61a	1	1	1	0	0	3	60		61b	1	1	1	1	1	1	6	100
	62a	1	1	1	0	1	4	80		62b	1	1	1	1	1	0	5	83
	63a	1	1	1	1	1	5	100		63b	1	1	1	1	1	0	5	83
	64a	1	0	0	0	0	1	20		64b	0	0	0	1	1	1	3	50
	65a	0	0	1	0	1	2	40		65b	1	1	1	0	1	1	5	83
	66a	1	0	1	1	0	3	60		66b	1	1	1	1	1	1	6	100
	67a	1	1	0	0	1	3	60		67b	0	0	1	1	1	1	4	67
	68a	1	1	1	0	0	3	60		68b	0	0	0	0	1	0	1	17
	69a	1	1	0	0	0	2	40		69b	0	0	0	1	1	0	2	33
	70a	1	0	1	1	1	4	80		70b	0	1	0	0	1	1	3	50
	71a	0	0	1	1	1	3	60		71b	1	1	0	1	1	1	5	83
	72a	0	0	0	0	0	0	0		72b	0	0	0	1	1	1	3	50
	73a	1	0	1	1	1	4	80		73b	0	1	0	1	1	1	4	67
		Average %						**56**			**Average %**							**70**

Note. Light grey shading indicates higher performance. Dark grey shading indicates average percentages. The computer-based version does not have an acceptable Cronbach's alpha value.

Table 41

Comparison of the Participants' Results on the Last Tasks in the German A2 Paper-Based and Computer-Based Tests

		Item No.										Item No.						
Test version	**Candidate No.**	**19**	**20**	**21**	**22**	**23**	**24**	**Σ**	**%**	**Test version**	**Candidate No.**	**20**	**22**	**23**	**24**	**25**	**Σ**	**%**
paper-based	74a	1	1	1	0	1	0	4	67	**computer-based**	74b	1	0	1	0	0	2	40
	75a	0	1	1	0	1	1	4	67		75b	1	1	1	1	1	5	100
	76a	1	1	1	0	1	0	4	67		76b	1	1	1	1	1	5	100
	77a	1	1	1	0	1	1	5	83		77b	1	0	1	1	1	4	80
	78a	0	1	1	1	1	1	5	83		78b	1	1	1	1	1	5	100
	79a	1	1	1	1	1	1	6	100		79b	1	1	1	1	1	5	100
	80a	1	1	1	0	1	1	5	83		80b	1	1	1	1	1	5	100
	81a	0	1	1	0	0	0	2	33		81b	1	1	1	1	1	5	100
	82a	0	1	1	0	1	1	4	67		82b	1	1	1	1	1	5	100
	83a	0	1	1	1	1	1	5	83		83b	1	1	1	1	1	5	100
	84a	1	1	1	0	1	1	5	83		84b	1	1	1	1	1	5	100
		Average %							**74**			**Average %**						**93**

Note. Light grey shading indicates higher performance. Dark grey shading indicates average percentages. Item 21 had to be deleted from the computer-based version in order to achieve an acceptable Cronbach's alpha value for the test.

Table 42

Comparison of the Participants' Results on the Last Tasks in the German B1 Paper-Based and Computer-Based Tests

		Item No.								Item No.						
Test version	**Candidate No.**	**26**	**27**	**28**	**29**	**Σ**	**%**	**Test version**	**Candidate No.**	**24**	**25**	**26**	**27**	**28**	**Σ**	**%**
paper-based	85a	0	0	1	1	2	50	**computer-based**	85b	1	0	0	1	0	2	40
	86a	0	0	1	1	2	50		86b	1	1	1	1	1	5	100
	87a	0	0	1	0	1	25		87b	1	0	0	1	0	2	40
	88a	0	0	1	1	2	50		88b	0	0	0	0	0	0	0
	89a	0	1	1	0	2	50		89b	1	0	1	0	0	2	40
	90a	0	1	0	1	2	50		90b	1	0	0	0	0	1	20
	91a	1	1	0	1	3	75		91b	1	1	0	1	1	4	80
	92a	1	1	0	0	2	50		92b	1	0	0	1	0	2	40
	93a	1	1	1	1	4	100		93b	1	0	1	0	1	3	60
	94a	0	1	0	0	1	25		94b	1	1	1	1	0	4	80
	95a	0	0	1	0	1	25		95b	1	1	1	1	0	4	80
	96a	1	0	1	0	2	50		96b	1	0	1	0	0	2	40
	97a	1	0	0	1	2	50		97b	1	0	1	0	1	3	60
	98a	0	1	0	0	1	25		98b	1	0	0	1	0	2	40
	99a	0	1	0	0	1	25		99b	1	0	0	0	1	2	40
	100a	0	1	1	0	2	50		100b	1	1	1	0	0	3	60
	101a	0	1	0	0	1	25		101b	1	1	0	1	0	3	60
	102a	0	1	0	0	1	25		102b	1	0	1	1	0	3	60
	103a	1	1	1	1	4	100		103b	1	1	1	0	0	3	60
		Average %					**47**			**Average %**						**53**

Note. Light grey shading indicates higher performance. Dark grey shading indicates average percentages. Item 25 had to be deleted from the paper-based version in order to achieve an acceptable Cronbach's alpha value for the test.

Table 43

Comparison of the Participants' Results on the Last Tasks in the German B2 Paper-Based and Computer-Based Tests

		Item No.									Item No.					
Test version	**Candidate No.**	**25**	**26**	**27**	**28**	**29**	**Σ**	**%**	**Test version**	**Candidate No.**	**23**	**24**	**25**	**26**	**Σ**	**%**
paper-based	104a	1	0	0	1	0	2	40	**computer-based**	104b	0	0	0	1	1	25
	105a	1	0	1	1	0	3	60		105b	0	0	0	0	0	0
	106a	0	1	0	1	1	3	60		106b	0	1	1	1	3	75
	107a	0	0	0	0	1	1	20		107b	0	1	1	1	3	75
	108a	0	0	0	1	1	2	40		108b	0	0	0	1	1	25
	109a	0	1	1	1	1	4	80		109b	0	1	0	0	1	25
	110a	0	1	1	1	1	4	80		110b	0	1	0	1	2	50
	111a	0	0	0	1	1	2	40		111b	0	0	1	1	2	50
	112a	0	1	0	1	1	3	60		112b	0	1	1	1	3	75
	113a	0	1	0	1	1	3	60		113b	0	0	1	1	2	50
	114a	1	0	0	1	1	3	60		114b	1	0	0	1	2	50
	115a	1	0	1	1	1	4	80		115b	0	0	0	0	0	0
	116a	1	1	1	1	1	5	100		116b	1	1	0	0	2	50
	117a	0	0	1	1	1	3	60		117b	0	1	1	1	3	75
	118a	1	1	1	1	1	5	100		118b	0	0	0	1	1	25
	119a	0	1	1	0	0	2	40		119b	1	1	1	1	4	100
	120a	1	1	1	0	0	3	60		120b	1	1	1	1	4	100
	121a	0	1	1	1	0	3	60		121b	0	1	0	0	1	25
	122a	1	1	1	1	1	5	100		122b	0	0	1	1	2	50
	123a	1	0	1	0	0	2	40		123b	0	1	0	1	2	50
	124a	1	0	0	1	0	2	40		124b	1	0	0	0	1	25
	125a	0	0	0	0	0	0	0		125b	1	0	1	0	2	50
	126a	1	1	1	1	1	5	100		126b	0	0	1	0	1	25
	127a	1	1	1	1	1	5	100		127b	0	1	1	0	2	50
		Average %						**62**			**Average %**					**49**

Note. Light grey shading indicates higher performance. Dark grey shading indicates average percentages. The paper-based version does not have an acceptable Cronbach's alpha value. Item 27 had to be deleted from the computer-based version in order to achieve an acceptable Cronbach's alpha value for the test.

Table 44

Comparison of the Participants' Results on the Last Tasks in the German C1 Paper-Based and Computer-Based Tests

		Item No.									Item No.							
Test version	**Candidate No.**	**25**	**26**	**27**	**28**	**29**	**Σ**	**%**	**Test version**	**Candidate No.**	**27**	**28**	**29**	**30**	**31**	**32**	**Σ**	**%**
paper-based	128a	1	1	1	0	1	4	80	**computer-based**	128b	1	1	1	1	1	0	5	83
	129a	0	0	1	1	0	2	40		129b	1	1	1	1	1	1	6	100
	130a	1	0	1	0	0	2	40		130b	1	0	0	1	1	0	3	50
	131a	0	1	1	0	1	3	60		131b	1	1	1	1	1	0	5	83
	132a	1	0	1	1	0	3	60		132b	1	0	1	0	1	0	3	50
	133a	1	1	1	0	1	4	80		133b	0	0	0	0	1	0	1	17
	134a	0	1	0	0	1	2	40		134b	1	1	1	1	1	1	6	100
	135a	1	0	1	0	1	3	60		135b	1	1	1	1	1	1	6	100
	136a	1	0	1	0	1	3	60		136b	1	1	1	0	1	1	5	83
	137a	0	0	1	0	1	2	40		137b	1	1	1	0	1	1	5	83
	138a	1	1	1	0	1	4	80		138b	0	1	1	0	0	1	3	50
	139a	1	0	1	1	1	4	80		139b	1	1	1	0	1	1	5	83
	140a	0	1	1	1	0	3	60		140b	0	0	1	1	1	0	3	50
		Average %						**60**			**Average %**							**72**

Note. Light grey shading indicates higher performance. Dark grey shading indicates average percentages.

When analysing the test results of the participants, there are three sets of tasks which cannot be considered in this analysis. The B1 English paper-based test, the English C1 computer-based test, and the B2 German paper-based test did not have a satisfactory Cronbach's alpha value, so they fail to measure the performance of the participants in a reliable way. For this reason, the present analysis cannot make any reliable claims related to the English B1 and C1, and the German B2 levels because even though the other version of the test at each of these language proficiency levels measures reliably, no reliable comparison could be made between them.

For the rest of the sets of tasks, the average percentage of the correct answers based on the answers of the participants were compared. As a result of the comparison, it seems that on the A2 and B2 English tests, the participants achieved a higher percentage on the paper-based tests; whereas, on the A2, B1, and C1 German tests they achieved higher scores on the computer-based tests. In the case of the A2 level English tests, the difference between the average percentage of the correct answers provided on the paper-based and on the computer-based version is 7%; whereas on the B2 level English test it is 10%. In the case of the A2 German test, the difference is 19%; on the B1 level, it is 6%; and on the C1 level, the difference is 12%. The results suggest varying degrees of difference between the different test versions. Disregarding those sets of tasks which did not have an acceptable Cronbach's alpha value, and only considering the ones, which did, it can be concluded that the participants generally did not have a lower performance on the audio-visual tasks than on the ATAO tasks.

Some limitations of the study, however, must also be taken into consideration. Firstly, these results might only apply to these specific tasks, and the possibility of different outcomes cannot be excluded with other similar audio-visual and ATAO tasks. As it can be seen from the results, in the English tests, the participants performed better on the ATAO

tasks than on the audio-visual ones at every language proficiency level. In contrast, in the German tests, the candidates' performance was higher on the audio-visual tasks than on the ATAO tasks on every proficiency level. As the ATAO tasks and the audio-visual tasks were designed for different recordings, these results could be just purely sample specific. A larger scale investigation with multiple audio-visual and ATAO task combinations should be carried out to arrive at generalizable results. Secondly, the errors of the computer programme recording the participants' answers cannot be excluded either. For instance, candidate no. 8 on the A2 level English test achieved 67% on the paper-based test but 0% on the computer-based version (Table 37). Similarly, candidate no. 49 on the B2 level English test achieved 100% on the paper-based test but 0% on the computer-based version (Table 39). If these results are only caused by computer data loss, they could severely distort the reliability of the analysis. As the data collection took place in the pilot phase of the language proficiency examination development project, such technical errors cannot be excluded.

5.3 Research question 3: Do the participants perceive the inclusion of audio-visual tasks as useful?

When considering supplementing the listening comprehension component of language proficiency tests, the opinion of the test-takers should also be taken into account. For this reason, the questionnaire the participants had to fill in after the tests also contained a construct referring to the usefulness of the audio-visual material for solving the tasks. In the case of the computer-based test where the last task was an audio-visual one, the questionnaire items inquired about the degree to which the participants felt that the video aided them in solving the last task. In contrast, in the paper-based version, there were no video aided tasks so the questionnaire items asked if the participants would have preferred to have video material while solving the tasks, and whether they would have found it useful. For the full questionnaires see Appendix 1C-2D. The tables (Table 45 - Table 53) below summarise the results of the questionnaires.

Table 45

Questionnaire Results: Necessity of the Video

Proficiency level	Language	Test version	Mean	Std. deviation
A2	English	Paper-based	3.68	0.96
		Computer-based	3.43	1.28
	German	Paper-based	2.89	0.82
		Computer-based	4.02	0.86
B1	English	Paper-based	3.10	1.17
		Computer-based	3.89	1.19
	German	Paper-based	2.97	1.17
		Computer-based	3.95	0.92
B2	English	Paper-based	2.71	1.29
		Computer-based	3.23	1.15
	German	Paper-based	3.52	1.07
		Computer-based	3.52	1.02
C1	English	Paper-based	3.04	0.93
		Computer-based	3.78	1.01
	German	Paper-based	2.69	1.25
		Computer-based	3.12	0.90

Regarding the paper-based version, the mean values of the answers seem to suggest medium to low preference for supplementing the listening component with audio-visual material, and they do not seem to think that it would help them in answering the questions. This result, however, could be slightly distorted by the fact that it was always the paper-based version of the test which was administered first, and the participants might not have had any previous experience with computer-based testing, let alone audio-visual tasks. Therefore, they might not have been able to imagine what it would mean to solve video aided listening tasks. This idea is further supported by the fact that the majority of the standard deviation values are relatively high, which indicates a high variance in the answers. Furthermore, in the case of the computer-based test questionnaire, the mean value at most language proficiency levels is above 3.50 and often close to 4.00, which indicates that the participants found the videos useful in solving the tasks.

Cases such as the A2 level German tests seem to also support the assumption that students might not have been able to imagine initially what solving audio-visual tasks is like

because the mean value for the paper-based version was $M = 2.89$ $SD = 0.82$, whereas for the computer-based version, it was $M = 4.02$, $SD = 0.86$. Similarly, on the B1 German paper-based version the mean value was $M = 2.97$ $SD = 1.17$, whereas for the computer-based version, it was $M = 3.95$, $SD = 0.92$. These two are the most prominent cases of the participants deeming the usefulness of the video differently across the different test versions; however, it can be observed that the mean values for the usefulness of the videos are always higher on the computer-based versions. Besides the lack of experience with such tasks, this phenomenon might also be explained by the nature of the platform and the possibility that the participants might more readily associate the digital platform with watching videos than the paper-based context.

To gain insight into the possible differences among the preferences of the different language proficiency levels regarding the usefulness and necessity of the videos, ANOVAs were calculated for each language and test versions. The analysis of variance (see Table 46) demonstrated that the difference among the mean values of the necessity of the video at the different language proficiency levels concerning the English paper-based tests was non-significant at the $p < .05$ level, $F(3,69) = 1.90$, $p = .14$. Post hoc analyses using the Duncan post hoc criterion indicated that the mean values were significantly lower for the B2 level test than for the A2 level test (see Table 47).

Table 46

One-Way Analysis of Variance of the Necessity of the Video Regarding the English Paper-Based Tests

Source	df	SS	MS	F	p
Between groups	3	7.39	2.46	1.90	.14
Within groups	69	89.53	1.30		
Total	72	96.91			

Note. $p < .05$

Table 47

Duncan Post Hoc Test for the Necessity of the Video Regarding the English Paper-Based Tests

Language Proficiency Level	N	Subset for α = 0.05	
		1	2
B2	26	2.71	
C1	17	3.04	3.04
B1	19	3.09	3.09
A2	11		3.68
p		.37	.13

The analysis of variance (see Table 48) demonstrated that the difference among the mean values of the necessity of the video at the different language proficiency levels concerning the English computer-based tests was non-significant at the $p < .05$ level, $F(3,69) = 1.50$, $p = .22$. Post hoc analyses using the Duncan post hoc analysis also indicated the same results (see Table 49).

Table 48

One-Way Analysis of Variance of the Necessity of the Video Regarding the English Computer-Based Tests

Source	df	SS	MS	F	p
Between groups	3	5.95	1.98	1.50	.22
Within groups	69	91.28	1.32		
Total	72	97.22			

Note. $p < .05$

Table 49

Duncan Post Hoc Test for the Necessity of the Video Regarding the English Computer-Based Tests

Language Proficiency Level	N	Subset for α = 0.05
B2	26	3.23
A2	11	3.43
C1	17	3.78
B1	19	3.89
p		.13

The analysis of variance (see Table 50) demonstrated that the difference among the mean values of the necessity of the video at the different language proficiency levels concerning the German paper-based tests was non-significant at the $p < .05$ level, $F(3,63) = 1.98$, $p = .13$. Post hoc analyses using the Duncan post hoc analysis also indicated the same results (see Table 51).

Table 50

One-Way Analysis of Variance of the Necessity of the Video Regarding the German Paper-Based Tests

Source	df	SS	MS	F	p
Between groups	3	7.22	2.41	1.98	.13
Within groups	63	76.54	1.26		
Total	66	83.76			

Note. $p < .05$

Table 51

Duncan Post Hoc Test for the Necessity of the Video Regarding the German Paper-Based Tests

Language Proficiency Level	N	Subset for α = 0.05
C1	13	2.69
A2	11	2.89
B1	19	2.97
B2	24	3.52
p		.06

The analysis of variance (see Table 52) demonstrated that the difference among the mean values of the necessity of the video at the different language proficiency levels concerning the German computer-based tests was non-significant at the $p < .05$ level, $F(3,63) = 2.73$, $p = .05$. Post hoc analyses using the Duncan post hoc criterion indicated that the mean values were significantly lower for the C1 level test than for the B1 and A2 level tests (see Table 53).

Table 52

One-Way Analysis of Variance of the Necessity of the Video Regarding the German Computer-Based Tests

Source	df	SS	MS	F	p
Between groups	3	7.32	2.44	2.73	.05
Within groups	63	56.32	.89		
Total	66	63.64			

Note. $p < .05$

Table 53

Duncan Post Hoc Test for the Necessity of the Video Regarding the German Computer-Based Tests

Language Proficiency Level	N	Subset for α = 0.05	
		1	2
C1	13	3.12	
B2	24	3.52	3.52
B1	19		3.95
A2	11		4.02
p		.24	.17

The analyses suggest that there are significant differences between the answers of the A2 and the B2 levels on the English paper-based test, and between the answers of the C1 and the B1 and A2 levels on the German computer-based test. This means that on the English paper-based test, the A2 level participants would have a significantly higher preference for having videos included next to the audio material than the B2 level participants. Furthermore, on the German computer-based tests, the A2 and B1 level participants found the videos accompanying the last tasks significantly more useful than the C1 level participants. These results seem to indicate that lower level test-takers might benefit from the presence of the videos more than those at a higher language proficiency level. However, because of the small sample size at each language proficiency level, this hypothesis should be further tested with larger samples. Nevertheless, the results of the questionnaires seem to indicate that the majority of the participants found the inclusion of videos non-disturbing and rather helpful in the listening component of the language tests.

6 Conclusion

The technological innovations of the 21st century have had substantial impact on the field of education. Recordings of lectures and online courses make education accessible for more and more people, and expose them to more and more audio-visual material as part of their education (Woolfitt, 2015). The field of foreign language teaching has recently also undergone a notable change caused by the increasing availability of audio-visual material for language learners. Using videos for foreign language learning purposes has become a frequently applied practice in foreign language classes (Suvorov, 2009), which can be assumed to have substantial influence on learning and practicing listening comprehension. As the aim of language testing is to assess a skill in an artificial situation which successfully emulates the intended real-life situation, the changes in teaching and using listening comprehension initiate the revision of how listening comprehension is measured in foreign language tests.

In addition, the present ways of testing listening comprehension were mostly developed in the 1980s with the emergence of the communicative language teaching. Based on Howe and Strauss (2007), this time marks a different generation than today's generation. Today's generation, namely *generation Z*, is exposed to audio-visual input much more frequently than to audio-only input, in contrast with previous generations, who had more exposure to audio-only input, for example, in the form of telephone conversations and radio broadcasts. These changes in the experience of the foreign language learners also necessitate the revision of the methods of testing listening comprehension.

Previous research does not provide unequivocal results related to the usefulness and the necessity of including audio-visual material into listening comprehension tests. Based on the findings of Bejar et al. (2000) and Ginther (2002), context-related and content-related visuals seem to enhance the comprehension of the aural input. In contrast, Ockey (2007) and

Londe (2009) found that the visual input had no effect on the performance of their participants. The contradictory results suggest that the issue needs to be further investigated. This is especially true for the Hungarian context, where at the time of conducting the present research, no research studies could be found in the topic of using audio-visual material in testing listening comprehension.

For these reasons, the aim of the present dissertation was to analyse whether including audio-visual tasks in the listening comprehension component of language examinations is necessary and desirable. In order to carry out this aim, the study investigated the following research questions:

1. Do the paper-based sets of tasks and the computer-based sets of tasks measure listening comprehension in an equally reliable way?
2. Does the performance of the test-takers on the audio-visual-to-audio-only tasks differ from their performance on the audio-visual tasks?
3. Do the participants perceive the inclusion of audio-visual tasks as useful?

The proposed issue was investigated in three phases. In the first phase, 16 sets of listening comprehension tasks were developed (8 English and 8 German) for four different language proficiency levels (i.e., A2, B1, B2, C1). Four sets of tasks for each language were developed to be administered in a paper-based format, and the other four were developed to be administered in a computer-based format. The paper-based sets of tasks were developed to be able to examine whether the computer-based sets of tasks measure listening comprehension as reliably as the paper-based sets of tasks. In order to be able to further investigate the possible effect of the audio-visual task on the participants' performance, the last task of each set was written for an audio-visual material, and in the case of the paper-based test, the visual input was removed during the test administration.

As the study also intended to investigate the participants' opinions about the usefulness and necessity of the audio-visual material, a questionnaire also had to be developed in the first data collection phase. Therefore, an interview schedule was developed, and 15 foreign language learners were asked to solve both the paper-based and the computer-based sets of tasks appropriate for their language proficiency level, and then share their opinion about the tasks in the form of a semi-structured interview. Based on the emerging themes of the interview and a relevant study in the topic found in the literature (i.e., Porsch et al., 2010), the first versions of the paper-based test questionnaire and the computer-based test questionnaire were developed. These versions were piloted in the second phase of the study with the help of four English learners, who were asked to solve the tasks, and then perform a think-aloud protocol on the questionnaires. Based on their feedback, the two versions of the questionnaire were finalised.

In the third data collection phase, 140 participants solved both the paper-based and the computer-based sets of tasks appropriate for their language proficiency levels. After solving each test, the participants were asked to fill in the questionnaire appropriate for the test version. The data collected from the participants was subjected to statistical analysis, such as ANOVA calculations, Cronbach's alpha analysis, item facility value calculations, and point-biserial correlation calculations.

Regarding the first research question, the results of the statistical analyses of the English and German paper-based and computer-based tests seem to indicate that the majority of the tests manage to measure the intended language proficiency levels in a satisfactory way for the present research purposes. The Cronbach's alpha values of the sets of tasks indicated that several tests had a satisfactory reliability value without any modifications. Such sets were the A2 and C1 English paper-based tests; the A2, B1, and B2 English computer-based tests; and the B1 and C1 German computer-based tests. Additionally, the

reliability of the B2 English paper-based test; the A2, B1, and C1 German paper-based tests; and the A2 and B2 German computer-based tests could be improved to be satisfactory by deleting some of the test items. There were also three sets of tasks with unsatisfactory reliability values which could not be improved by eliminating any items. These were the B1 level English paper-based test, the C1 level English computer-based test, and the B2 level German paper-based test.

In addition to the test results, the participants' answers to the questionnaire constructs named *disturbing features of the test, structure of the test,* and *perceived difficulty of the test* were also subjected to analysis. The mean values and the standard deviations of the constructs suggest that the participants at all proficiency levels agreed that they did not perceive any major disturbing factors in the test. The majority of the participants also agreed that the test is well-structured. Regarding the perceived difficulty of the test, only the B2 German paper-based test was indicated to be rather difficult; regarding the rest of the tests, the participants indicated moderate to low difficulty. The combined results of the test and the questionnaire data seem to suggest that the overall results are not worse on the computer-based than on the paper-based tests so the computer-based sets of tasks do not measure participants' listening skills in a less reliable way than the paper-based sets of tasks.

Regarding the second research question, the participants' performance on the last task was examined both on the paper-based and on the computer-based tests. The results indicate that on the A2 and B2 English tests, the participants achieved a higher percentage on the paper-based tests; whereas, on the A2, B1, and C1 German tests they achieve higher scores on the computer-based tests. In the case of the A2 level English tests, the difference between the average percentage of the correct answers provided on the paper-based and on the computer-based version is 7%; whereas on the B2 level English test it is 10%. In the case of the A2 German test, the difference is 19%; on the B1 level, it is 6%; and on the C1

level, the difference is 12%. The results suggest varying degrees of difference between the different test versions. However, it can be concluded that the audio-visual tasks do not measure listening comprehension less reliably than the audio-only tasks.

In connection with the third research question, which enquired about the participants' opinions on the usefulness and necessity of the audio-visual tasks, the results suggest that test-takers with lower language proficiency levels seem to benefit more from the presence of the visual input than the test-takers with higher language proficiency levels. The ANOVA calculations indicate that the majority of the participants found the inclusion of videos non-disturbing and rather helpful in the tests.

In conclusion, the results of the present study appear to indicate that the computer-based sets of tasks do not measure participants' listening skills in a less reliable way than the paper-based sets of tasks. Furthermore, the comparison of the last tasks of the paper-based and the computer-based sets of tasks also indicates that the participants' performance on the audio-visual tasks was similar — in terms of the test scores — to the ATAO tasks. Thirdly, the majority of the participants found the presence of the videos in the audio-visual tasks non-disturbing or even helpful, which seems to be in accordance with the findings of Bejar et al. (2000) and Ginther (2002). The popularity of consuming audio-visual materials in people's everyday life appears to indicate that the criterion-related validity of listening comprehension tests could be improved by the inclusion of audio-visual tasks because it would raise the authenticity of the test. Additionally, the results of the present dissertation show that the reliability of such a test would not be lower than that of a traditional audio-only listening test, and that the participants do not seem to find the presence of the audio-visual material disturbing. In fact, many of the lower language proficiency level participants found it helpful in solving the task. For this reason, the revision of the listening

comprehension component of language examinations can be proposed, and its supplementation with audio-visual material could be encouraged.

7 Limitations of the study and implications for further research

As any research endeavour, the present study also has some limitations which should be addressed. First, related to the data collection procedures, the main limitation is that all the participants of the third phase had to first execute the paper-based tests and only then the computer-based version. Therefore, when answering the questionnaire items related to the necessity of the video, more specifically, if a video could have aided them in solving the tasks, many of the participants might not have been able to accurately imagine what such a video-aided listening task would have been like. For this reason, some participants' answers to these questions might not be fully reliable because of the order effects. The idea that some of the participants were not able to accurately imagine the influence of a video accompanying a listening task is also supported by some of the results of the study, which suggest that in the case of the German A2 and C1 tests, the participants managed to achieve a higher score on the computer-based test than on the paper-based test; and in the case of the computer-based test the majority of the participants found the video useful in solving the task even though in the paper-based test they deemed the addition of a video unnecessary. Therefore, repeating the data collection in a way that the paper- and the computer-based tests are administered in a counterbalanced design among the participants (i.e., half of the participants write the computer-based test first and the paper-based test second, and the other half of the participants write the paper-based test first and the computer-based second). As an alternative, administering the two test versions, and the questionnaire items related to the necessity and usefulness of the video among participants who are already familiar with solving audio-visual tasks might be able to yield more reliable results. However, this might not be feasible in the current Hungarian context because using audio-visual material and audio-visual listening comprehension tasks are probably not part of most students' learning experience.

In addition, the reliability of the questionnaire results could be further enhanced, especially in the case of the paper-based test questionnaire by administering separate questionnaires after each task and not only at the end of the full listening comprehension component. In in this way, the questionnaire could have caught the differences related to the necessity of a video regarding the different text types and item formats. As the data collection for the present dissertation was carried out as part of a larger project, such decisions could not be controlled by the author of this dissertation.

Another limitation might emerge in connection with the data collection platform of the computer-based test. As discussed in the data collection section of the dissertation, at the time of the data collection the digital platform was still in the development phase. For this reason, technical issues related to the recording of the participants' answers could not be completely avoided. Because of such technical difficulties, the loss of some research data was inevitable. To ensure the reliability of the results, in cases where the digital platform failed to record all the data provided by a certain participant, those participants' results were completely eliminated from the data pool. Furthermore, the technical difficulties caused stress for some of the participants who decided to opt out of the data collection completely because of them.

Another issue related to the data collection platform being in the development phase is that it was unable to collect metadata about the participants' task-solving processes, for example, logging the amount of time test-takers spent on a single task, logging whether the test-takers paused the video recording and at which point of the recording they paused it, and how many times they paused it. Such information could have provided further insights into the participants' task-solving processes and it could have provided information regarding the discourse features of the different text types.

In connection with the limitations of the results, the usefulness and necessity of the videos should also be assessed carefully. In the current study, the results suggest that the majority of the participants found the video useful for solving the audio-visual task. However, based on the collected data, this result cannot necessarily be generalised to other item formats. Based on the results of the present study, it can only be claimed that the videos appear to be useful for a particular text type with a particular item format. Therefore, further research would be necessary featuring more audio-visual tasks written for a variety of text types with a variety of different item formats.

Additionally, the size of the data also does not enable the generalisation of the results. Even though the total number of participants is 140, there were not enough participants for each language proficiency level to be able to provide generalisable data. To remedy this deficiency, a large-scale study should be conducted with larger sample sizes for each language proficiency level. However, such an endeavour might face obstacles in connection with data collection on the A2 and the B1 level language proficiencies because, according to NYAK (2019), language examinations on the A2 and the B1 language proficiency levels are not as frequently attended by language learners in Hungary as the B2 and C1 level language proficiency examinations. Taking A2 level language examinations in Hungary is especially infrequent because currently it does not provide an accredited language certificate (NYAK, 2019).

8 Pedagogical implications

Using audio-visual materials in language examinations can have several beneficial effects on the field of language testing. On the basis of the conclusions of the present study, including audio-visual tasks into the listening comprehension component of language examinations can be methodologically supported, and it can enhance the criterion-related validity of the test. In addition, the inclusion of the audio-visual material might be able to improve the performance of the test-taker. According to statistical analyses using the many-facet Rasch measurement approach conducted by G. Dávid (personal communication, February 3, 2020), it appears that audio-visual tasks help students achieve better performance on the listening component of the test developed in the project. Furthermore, the inclusion of the audio-visual material would also encourage language examination centres to move their examinations to a computer-based platform. A well-built digital platform could provide new benefits for all stakeholders, for example, such a platform would be able to track the test-taking strategies of the candidates, or metadata about their task solving processes; thus, providing valuable insights both for the language testing and the foreign language education community.

Besides its positive effect on the field of language testing, supplementing the listening comprehension component of foreign language examinations with audio-visual tasks could possibly have positive washback effect on foreign language education in the Hungarian context. One of the possible positive effects using audio-visual material in foreign language examinations can result in is the increased use of video material in foreign language classrooms. If students want to practice for a foreign language examination containing audio-visual tasks, the teachers and tutors might also begin to develop audio-visual practice tasks for their classes. Such a practice is already present in today's

language classrooms, but it should be more structured and more aligned with the intended learning outcomes.

Furthermore, if using audio-visual material becomes a regular part of foreign language education, it might also start to influence other fields of education in Hungary, and the use of audio-visual materials could become part of the regular classroom activities. Besides the audio-visual material produced by the teachers for educational and practice purposes, the students could also get engaged in creating their own audio-visual material. On the one hand, these could be used as part of the classroom assessment; and on the other hand, it would teach the students valuable new digital skills which are becoming essential in most professional careers, and which could also facilitate the development of the students' autonomous learning skills. Such practices have already started to gain traction and to be successfully implemented in the Dutch education system. According to Woolfitt (2015), making video recordings of university lectures available for students can aid them during their examination preparations, and assignments requiring them to video record and upload their presentations can greatly improve students' presentation skills. Even though Woolfitt's (2015) study discusses the uses of audio-visual material in tertiary education, the same practices could probably be successfully adapted for lower levels of education as well.

When introducing such a major innovation into the field of education and testing, the background knowledge of the teachers also needs to be taken into consideration (Woolfitt, 2015). The use of audio-visual materials in foreign language classrooms in Hungary is a highly under-researched area so it can be presupposed that the majority of current foreign language teachers in Hungary are not extensively familiar or maybe not comfortable with creating and/or using audio-visual materials in their classes. For this reason, this pedagogical shift should begin at the level of teacher training by providing instructions in the methodological and technological background of creating and using audio-visual materials

in education for teacher trainees, and by enabling already practicing teachers to access the same types of opportunities.

Furthermore, introducing audio-visual material as a regular part of classroom activity necessitates the availability of a certain amount of technological equipment in the Hungarian educational institutions. Such technological equipment could contain, for instance, smartboards, projectors, tablets and Internet access for every student during class time. It must be acknowledged that this would probably require a large financial investment. Nevertheless, such an improvement could help the students acquire valuable knowledge and skills they can later benefit from in their careers. In addition, students using smartphones in class is currently a highly debated issue (Anshari, Almunawar, Shahril, Wicaksono & Huda, 2017), and incorporating the use of smartphones to watch and create audio-visual material as part of the class work might reduce their disturbing effects of students' attention levels.

Besides the positive washback effects, introducing the regular use of audio-visual tasks into foreign language teaching and testing could also result in a negative washback. For instance, teachers might start using videos in their classes without any sound methodological reason for the sake of "fashion", which would then result in filibustering the learning process, instead of using audio-visual materials which seem to be reasonable and suitable to be incorporated into the particular learning material because of their potential of providing a more vivid learning experience. Videos should only be used to enhance students' understanding of the material, and only at places where it is methodologically justifiable. However, applied under the right terms, using audio-visual materials in class can have notable positive effects on the learning process and the learning environment.

In conclusion, it can be argued that if language testing resists to include new task types — however unorthodox they may seem at present —, language proficiency tests risk becoming obsolete and not being able to authentically test real life language use. For this

reason, even if stakeholders feel reluctant to invest resources into the technological assets necessary for making audio-visual materials more prominent parts of language testing and education, a change of perspective about the use of technology in foreign language testing and education would be in order. Computer-based tests would not only be modern because of the computer use itself, but also because of the ways and new approaches a well-built digital platform could provide for the stakeholders. In addition, making audio-visual tasks a part of language examinations can be expected to have several beneficial effects on foreign language teaching and education in general as well. Further studies in the field could lead to a change of perspective about the use of technology in foreign language education and foreign language testing, and it might result in a greater beneficence from the contemporary technological developments in these fields.

9 Feasibility issues

Based on the discussion presented in the conclusion section, it appears that supplementing the listening comprehension component of foreign language examinations with audio-visual tasks would be necessary and desirable in order to maximise the criterion-related validity of the test. In addition, the results also suggest that there seem to be no methodological obstacles to this initiative because the majority of the sets of tasks containing ATAO tasks and sets of tasks also containing an audio-visual task measured listening comprehension similarly. Nevertheless, besides the reliability and validity issues, concerns related to feasibility should also be addressed.

The main feasibility concerns are related to the financial aspect of introducing the audio-visual tasks to the listening comprehension component of language examinations. First, because of the possible new challenges that writing items for audio-visual tasks can impose on the item writers and actors, it is likely that the item writers would need to be specially trained. Administering such training would naturally require some financial investment. An even larger investment would be necessary for the currently existing language examination centres to be able to shift their examinations from a paper-based platform to a digital one. This would require the purchase of such technological equipment which is currently not available at most language examination centres in Hungary. In addition, the development of audio-visual tasks also necessitates financial investment because the copyright of the videos used for task development has to be purchased. Paying for the copyright is unavoidable in this situation because using the videos in language examinations counts as a for profit use of the material. Besides the possibly high expenses, there are other issues that can emerge in connection with purchasing the copyright of videos. First, the owner of the video might not be willing to agree to sell the copyright. Second, even if they are willing to sell them, the copyright might be excessively expensive. Even if the

language examination centre is willing to invest the necessary amount of money, the video might prove not to be suitable for item writing purposes in the end. Just as in the case of any data collection instrument, the results of the pilot might suggest that the developed audio-visual task is not measuring the intended language proficiency level appropriately, and the items of the task might need to be re-worked. In some cases, improving the reliability of the items might not even be possible, and the task would need to be completely discarded; thus, it would not be able to return the copyright investment.

A possible way to avoid the emerging copyright issues is for the language examination centres to create their own audio-visual materials to use for task development. However, as the CEFR descriptors (Council of Europe, 2001, 2018) appoint the native speaker as the standard, which foreign language examinees have to adhere to, the audio-visual material used for task development should feature interactions among native speakers. This would require the language examination centres to hire native speaker actors to record the audio-visual material. In the Hungarian context, this would probably not be feasible, and even if native speaker actors could be found, the expenses would likely rival the price of purchasing the copyright of an already existing video.

One way to overcome the above-mentioned problems and to reduce the costs of audio-visual task development would be to not restrict the used audio-visual material to those which are produced by native speakers only. Also using content featuring non-native speakers would enable the item writers to select the material from a larger data pool, so they would be more likely able to select audio-visual content with lower copyright prices. The practice of using content produced by non-natives speakers who have a high language proficiency in the foreign language in question is not a new practice. Language examinations such as the *Cambridge Assessment English* (Cambridge Assessment, 2019) already use non-native speaker material in their listening comprehension examination tasks (e.g.,

Business Vantage and Business Higher). Furthermore, in the case of the English language in the Hungarian context, the majority of the examinees are more likely to encounter situations where they have to speak English with other non-native speakers during their career or studies. For this reason, using non-native audio-visual material in foreign language examinations, especially in the case of the English language, appears to be a reasonable decision.

In conclusion, supplementing the listening comprehension component of foreign language examinations with an audio-visual component is worth serious consideration. The results of the present study suggest that doing so would improve the criterion-related validity of the test. However, there are several difficulties which should be addressed in connection with such an innovation and addressing these issues would probably require a change of perspective in the Hungarian language testing community, and it would necessitate a considerable financial investment from the part of the language examination centres. Nevertheless, because of the constant technological development, more and more leading language examination centres in the world could be expected to shift their examinations to a digital platform in the near future; thus, such a change will probably also be inevitable in the Hungarian context.

References

Akkreditációs kézikönyv [Accreditation Manual]. (2018). Retrieved August 22, 2019 from https://nyak.oh.gov.hu/nyat/doc/ak2018/ak2018.htm

Anshari, M., Almunawar, N. M., Shahril, M., Wicaksono, K. D., & Huda, M. (2017). Smartphones usage in the classrooms: Learning aid or interference? *Education and Information Technologies, 22*(6), 3063–3079.

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

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

Bachman, L. F., & Palmer, A. S. (2010). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative language ability. *TESOL Quarterly, 16*(4), 449–465.

Bárdos, J. (2005). *Élő nyelvtanítás-történet*. Budapest, HU: Nemzeti Tankönyvkiadó.

Bates, A. W. (2015). *Teaching in a digital age: Guidelines for designing teaching and learning for a digital age* [PDF file]. Retrieved August 22, 2019 from https://teachonline.ca/sites/default/files/pdfs/teaching-in-a-digital-age_2016.pdf

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper*. Princeton, NJ: Educational Testing Service.

Berg, T. (1987). The case against accommodation: Evidence from German speech error data. *Journal of Memory and Language, 26*, 277–299.

Berne, J. E. (1995). How does varying pre-listening activities affect second language listening comprehension? *Hispania, 78*(2), 316–329.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071.

Brazil, D. (1983). Intonation and discourse: Some principles and procedures. *Text, 3*(1), 39–70.

Bregman, A. (1978). The formation of auditory streams. In J. Requin (Ed.), *Attention and performance* (pp. 63–75). Hillsdale, NJ: Earlbaum.

Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics, 18*, 171–191.

Bowles, M. (2010). *The think-aloud protocols controversy in second language research*. New York, NY: Routledge.

Brown, G. (1986). Investigating listening comprehension in context. *Applied Linguistics, 7*, 284–302.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge, UK: Cambridge University Press.

Brynjolfsson, E., & Mcafee, A. (2014). *The second machine age*. New York, NY: W.W. Norton & Company.

Buck, G. (1991). The test of listening comprehension: An introspective study. *Language Testing, 8*(1), 67–91.

Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing, 11*(2), 145–170.

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

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119–157.

Burgoon, J. (1994). Non-verbal signals. In M. Knapp, & G. Miller (Eds.), *Handbook of interpersonal communication* (pp. 344–393). London, UK: Routledge.

Cambridge Assessment (2019). Cambridge English computer-based exams. Retrieved August 22, 2019 from https://www.cambridge-exams.ch/exams/CB_exams.php

Canale, M. (1983a). From communicative competence to communicative language pedagogy. In J. C. Richards, & R. W. Schmidt (Eds.), *Language and communication* (pp. 2–27). London, UK: Longman.

Canale, M. (1983b). On some dimensions of language proficiency. In J. W. Jr. Oller (Ed.), *Issues in language testing research* (pp. 333–342). Rowley, MA: Newbury House.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.

Carr, T., & Duncan, J. (1987). The VCR revolution: feature films for language and cultural proficiency. In D. W. Birckbichler (Ed.), *Proficiency, policy, and professionalism in foreign language education* (pp. 92–105). Lincolnwood, IL: National Textbook Co.

Carroll, J. B. (1968). The psychology of language testing. In A. Davies (Ed.), *Language testing symposium: A psycholinguistic approach* (pp. 46–69). London, UK: Oxford University Press.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995) Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics*, *2*, 5–35.

Cervantes, R., & Gainer, G. (1992). The effects of syntactic simplification and repetition on listening comprehension. *TESOL Quarterly*, *26*(4), 767–770.

Chafe, W. (1980). The deployment of consciousness in the production of a narrative. In W. Chafe (Ed.), *The pear stories* (pp. 9–50). Norwood, NJ: Ablex.

Chafe, W. (1982). Integration and involvement in speaking, writing and oral literature. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 35–53). Norwood, NJ: Ablex.

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

Chapelle, C., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge, UK: Cambridge University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: M.I.T. Press.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York, NY: Harper and Row.

Clark, J. L. D. (1972). *Foreign language testing: Theory and practice.* Philadelphia, PA: Center for Curriculum Development.

Clark, H. H., & Clark, E. V. (1977). *Psychology and language: An introduction to psycholinguistics*. New York, NY: Harcourt Brace Jovanovich.

Cohen, A. D. (1980). *Testing language ability in the classroom.* Rowley, MA: Newbury House.

Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, 3(4), 307–331.

Cohen, A. D. (2011). *Strategies in learning and using a second language* (2nd ed.). Harlow, UK: Pearson Education Limited.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education*. (6th ed.). New York, NY: Routledge Falmer.

Combi, C. (2015). *Generation Z: Their voices, their lives*. London, UK: Hutchinson.

Conrad, L. (1985). Semantic versus syntactic cues in listening comprehension. *Studies in Second Language Acquisition, 7*, 59–72.

Conrad, L. (1989). The effects of time-compressed speech on listening comprehension. *Studies in Second Language Acquisition*, *11*, 1–16.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

Council of Europe (2018). *The CEFR Companion Volume with New Descriptors*. Retrieved August 22, 2019 from: https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Creswell, J. W. (2009). *Research design: Qualitative, quantitative and mixed methods approaches* (3rd ed.). London, UK: Sage.

Demyankov, V. Z. (1983). Understanding as an interpreting activity. *Voprosy Yazykoznaniya*, *32*, 58–67.

De Boer, J. (2013). *Learning from video: Viewing behavior of students*. Enschede, NL: Ipskamp Drukkers B.V.

De Vera, J. M., & McDonnell, J. (1985). Video: A media revolution? *Communication Research Trends*, *6*(2), 1–8.

Dunkel, P. (1991). Computerized testing of nonparticipatory L2 listening comprehension proficiency: An ESL prototype development effort. *Modern Language Journal*, *75*, 64–73.

Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal*, *77*(2), 180–191.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

ETS TOEFL (2019). About the TOEFL iBT Test. Retrieved August 22, 2019 from https://www.ets.org/toefl/ibt/about

Field, J. (2009). *Listening in the language classroom*. Cambridge, UK: Cambridge University Press.

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

Folley, S. (2015). The effect of visual cues in listening comprehension: Pedagogical implications for non-native speakers of English. *Culminating Projects in English. Paper 39*. Retrieved August 22, 2019 from https://repository.stcloudstate.edu/engl_etds/39/

Fortune, A. J. (2004). *Testing listening comprehension in a foreign language: Does the number of times a text is heard affect performance*? (Unpublished master's thesis). University of Lancaster, Lancaster.

Fransen, J. (2015). *Instrumentatie van beteknisvolle interacties*. Den Haag, NL: Inhholland.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London, UK: Routledge.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing*, *19*, 133–167.

Glasser, R. (1991). Expertise and assessment. In M. C. Wittrock, & E. L. Baker (Eds.), *Testing and cognition* (pp. 17–30). Englewood Cliffs, NJ: Prentice Hall.

Gósy, M. (2000). *A hallástól a tanulásig*. Budapest, HU: Nikol.

Greenberg, A. D., & Zanetis, J. (2012). *The impact of broadcast and streaming video in education* [PDF file]. Retrieved August 22, 2019 from http://www.cisco.com/web/strategy/docs/education/ciscovideowp.pdf

Guo, P. J., Kim, J., & Rubin, R. (2014). *How video production affects student engagement: An empirical study of MOOC videos* [PDF file]. Retrieved August 22, 2019 from http://pgbovine.net/publications/edX-MOOC-video-production-and-engagement_LAS-2014.pdf

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). New York, NY: Macmillan.

Hansch, A., Newman, C., Hillers, L., Shildhauer, T., McConachie, K., & Schmidt, P. (2015). *Video and online learning: Critical reflections and findings from the field* [PDF file]. Retrieved August 22, 2019 from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2577882

Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 241–268). New York, NY: Cambridge University Press.

Howe, N., & Strauss, W. (2007). The next 20 years: How customer and workforce attitudes will evolve. *Harvard Business Review*, *85*(7–8), 41–52.

Hymes, D. H. (1971). *On communicative competence*. Philadelphia, PA: University of Pennsylvania Press.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride, & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Harmondsworth, UK: Penguin.

Iimura, H. (2007). The listening process: Effects of question types and repetition. *Language Education and Technology, 44*, 75–85.

Johnson, L., Adams Becker, S., Estrada, V., & Freeman, A. (2015). *NMC horizon report: 2015 higher education edition*. Retrieved August 22, 2019 from http://cdn.nmc.org/media/2015-nmc-horizon-report-HE-EN.pdf

Kaltura Report (2015). *The state of video in education – 2015: A Kaltura Report* [PDF file]. Retrieved August 22, 2019 from https://corp.kaltura.com/

Kaltura Report (2019). *The state of video in education – 2019: Insights and trends* [PDF file]. Retrieved August 22, 2019 from https://corp.kaltura.com/

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*, 135–170.

Kellerman, S. (1990). Lip service: The contribution of the visual modality to speech perception and its relevance to the teaching and testing of foreign language listening comprehension. *Applied Linguistics*, *11*, 272–280.

Kelly, P. (1991). Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners. *IRAL, 29*, 135–149.

Kirschner, P. A., & van Merriënboer, J. J. G. (2013). Do learners really know best? Urban legends in education. *Educational Psychologist*, *48*(3), 169–183. doi:10.1080/00461520.2013.804395

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London, UK: Routledge.

Kreckel, M. (1981). Tone units as message blocks in natural discourse: Segmentation of face-to-face interaction by naïve native speakers. *Journal of Pragmatics*, *5*, 459–476.

Kuang-yun, T. (2007). *Teaching English using the internet and the multiple intelligences approach*. Retrieved August 22, 2019 from https://books.google.hu/books?id=eKLJNSzYzKMC&printsec=frontcover&hl=hu&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Kubanyiova, M. (2015). Ethics in research. In J. D. Brown, & C. Coombe (Eds.), *The Cambridge guide to research in language teaching and research* (pp. 176–182). Cambridge, UK: Cambridge University Press.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London, UK: Longman.

Larson, J. W. (1989). S-CAPE; A Spanish Computerized Adaptive Placement Exam. In W. F. Smith (Ed.), *Modern technology in foreign language education: Applications and projects* (pp. 277–289). Skokie, IL: National Textbook Company.

Londe, Z. C. (2009). The effects of video media in English as a second language listening comprehension tests. *Issues in Applied Linguistics*, *17*(1), 41–50.

Lonergan, J. (1984). Video in language teaching. Cambridge, UK: Cambridge University Press.

Lund, R. (1991). A comparison of second language listening and reading comprehension. *Modern Language Journal*, *75*(2), 196–204.

MacWilliam, I., (1986). Video and language comprehension. *ELT Journal*, *40*, 131–135.

Marslen-Wilson, W., & Tyler, L., (1980). The temporal structure of spoken language comprehension. *Cognition*, *8*, 1–72.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 43–52. doi:10.1207/S15326985EP3801_6

Maykut, P., & Morehouse, R. (2002). *Beginning qualitative research: A philosophic and practical guide*. London, UK: The Falmer Press.

McCracken, G. (1988). *The long interview*. Newbury Park, CA: SAGE Publications.

McCrindle, M., & Wolfinger, E. (2014). *The ABC of XYZ: Understanding the global generations* (3rd ed.). Bella Vista, AU: McCrindle Research.

McNamara, T. F. (1996). *Measuring second language performance*. London, UK: Addison Wesley Longman.

McNamara, T. (2000). *Language testing*. Oxford, UK: Oxford University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York, NY:American Council on Education.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*(9), 741–749.

Meunier, L. E. (1994). Computer adaptive language tests (CALT) offer a great potential for functional testing. Yet, why don't they? *CALICO Journal*, *11*(4), 23–39.

Morrow, K. (1977). *Techniques of evaluation for a notional syllabus*. London, UK: Royal Society of Arts.

Morrow, K. (1979). Communicative language testing: Revolution or evolution? In C. J. Brumfit, & K. Johnson (Eds.), *The communicative approach to language teaching* (pp. 143–157). Oxford, UK: Oxford University Press.

Noijons, J. (1994). Testing computer assisted language tests: Towards a checklist for CALT. *CALICO Journal*, *12*(1), 37–58.

Nyelvvizsgáztatási Akkreditációs Központ (NYAK) [Educational Authority Accreditation Centre for Foreign Language Examinations]. (2019). Nyelvvizsga-statisztikák (vizsgázók száma alapján): Évek szerinti bontás szintek szerint — 2009-2018 [Foreign language examination statistics (on the basis of the number of test-takers): Annual statistics by language proficiency levels — 2009-2018]. Retrieved August 22, 2019 from https://nyak.oh.gov.hu/doc/statisztika.asp?strId=_07

Ockey, G. (2007). Construct implication of including still image or video in computer-based listening tests. *Language Testing*, *24*, 517–537.

Oller, J. W. Jr. (1976). Evidence of a general language proficiency factor. *Die Neueren Sprachen*, *76*, 165–174.

OM Rendelet [Education Decree]. (2006). *15/2006. (IV. 3.) OM rendelet az alap- és mesterképzési szakok képzési és kimeneti követelményeiről* [*Education Decree No. 15/2006 (IV. 3.) on the Training and Outcome Requirements to Undergraduate Degrees (B.A./B.Sc.) and Graduate Degrees (M.A./M.Sc.)*]. Retrieved August 22, 2019 from http://cdn.felvi.hu/bin/content/dload/jogszabalyok/15_2006_alap_mester_kkk_20080201.pdf

Otsuka, K. (2004, August). *How to determine the optimal number of listening opportunities for listening comprehension tests among Japanese high school learners of English*? Paper presented at the 9th Conference of the Pan-Pacific Association of Applied Linguistics, Chonan, Republic of Korea.

Palfrey, J., & Gasser, U. (2008). *Born digital: Understanding the first generation of digital natives*. New York, NY: Basic Books.

Petőné Honvári, J. (2014). Hallás és olvasás utáni szövegértés különbségeinek vizsgálata egy negyedikes osztályban. *Új Pedagógiai Szemle*, *9–10*, 58–68.

Porsch, R., Grotjahn, R., & Tesch, B. (2010). Hörverstehen und Hör-Sehverstehen in der Fremdsprache – unterschiedliche Konstrukte? *Zeitschrift für Fremdsprachenforschung, 21*(2), pp. 143–189.

Progosh, D. (1996). Using video for listening assessment: Opinions of test-takers. *TESL Canada Journal*, *14*(1), 34–44.

Raffler-Engel, W. (1980). Kinesics and paralinguistics: A neglected factor in second language research and teaching. *Canadian Modern Language Review, 36*, 225–237.

Rasch, G. (1960). *Probabilistic models for some intelligence attainment tests*. Chicago, IL: University of Chicago Press.

Richards, J. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly*, *17*(2), 219–240.

Rivers, W. M. (1981). *Teaching foreign-language skills*. Chicago, IL: The University of Chicago Press.

Rost, M. (1990). *Listening in language learning.* New York, NY: Longman.

Rubin, J. (1995). An overview to 'A guide for the teaching of second language listening'. In D. J. Mendelsohn, & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 7–11). San Diego, CA: Dominie Press, Inc.

Sakai, H. (2009). Effect of repetition of exposure and proficiency level in L2 listening tests. *TESOL Quarterly*, *43*(2), 360–372.

Sheskin, J. D. (2011). *Handbook of parametric and nonparametric statistical procedure* (5th ed.). Boca Raton, FL: CRC Press.

Siemens, G., Gašević, D., & Dawson, S. (2015). *Preparing for the digital university: A review of the history and current state of distance, blended, and online learning* [PDF file]. Retrieved August 22, 2019 from http://linkresearchlab.org/PreparingDigitalUniversity.pdf

Simons, R., & Bolhuis, S. (2004). Constructivist learning theories and complex learning environments. *Oxford Studies in Comparative Education*, *13*(1), 13–25.

Strauss, W., & Howe, N. (1997). *The fourth turning: An American prophecy*. New York, NY: Broadway Books.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question-type. *Language Testing*, *8*, 23–40.

Sueyoshi, A., & Hardison, D. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*, 661–699.

Sulaiman, M. Z., & Kahn, A. M. (2019). Computer assisted language testing (CALT): Issues and challenges. *International Journal of Higher Education and Research*, *9*(1), 1–11.

Suvorov, R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. In C. A. Chapelle, H. G. Jun, & I. Katz (Eds.), *Developing and evaluating language learning materials* (pp. 53–68). Ames, IA: Iowa State University.

Suvorov, R. (2011). The effects of context visuals on L2 listening comprehension. *University of Cambridge ESOL Examinations Research Notes*, *45*, 2–8.

Szabó, G., & Nikolov, M. (2013). An analysis of young learners' feedback on diagnostic listening comprehension tests. In J. Mihaljević Djigunović, & M. Medved Krajnović (Eds.), *UZRT 2012: Empirical studies in English applied linguistics* (pp. 7–21). Zagreb, HR: FF Press.

Valette, R. (1977). *Modern language testing* (2nd ed.). San Diego, CA: Harcourt Brace Jovanovich.

Wagner, E. (2002). Video listening tests: A pilot study. *Working Papers in TESOL & Applied Linguistics, Teachers College, Columbia University*, *2*(1), 1–39.

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology*, *11*(1), 67–86.

Wagner, E. (2014). Using Unscripted Spoken Texts in the Teaching of Second Language Listening. *TESOL Journal*, *5*(2), 288–311.

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls and unexpected consequences of implementing large-scale computerized testing. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computer adaptive testing: A primer* (2nd ed., pp. 271–299). Mahwah, NJ: Erlbaum.

Weir, C. (1993). *Understanding and developing language tests*. Hemel Hempstead, UK: Prentice-Hall.

Weir, C. J. (1983). The Associated Examining Board's test of English for academic purposes: An exercise in content validation events. In A. Hughes, & D. Porter (Eds.), *Current developments in language testing* (pp. 147–153). London, UK: Academic Press.

Woolfitt, Z. (2015). *The effective use of video in higher education* [PDF file]. Retrieved August 22, 2019 from https://www.inholland.nl/media/10230/the-effective-use-of-video-in-higher-education-woolfitt-october-2015.pdf

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.

Yousef, A. M. F., Chatti, M. A., & Schroeder, U. (2014, March). *Video-based learning: A critical analysis of the research published in 2003-2013 and future visions*. Paper presented at the eLmL 2014: The Sixth International Conference on Mobile, Hybrid and On-line Learning, Barcelona, Spain.

Appendices

Appendix 1A – Table A: Main characteristics of the A2 English language paper-based task set

Table A

Main characteristics of the A2 English language paper-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Global understanding and understanding one piece of specific information	Short airport announcements and instructions (approx. 2.5 minutes/1 recording; altogether 3 recordings)	1 true or false item for the main message per recording 1 multiple choice question (3 options) for a specific detail per recording	6
2	Listening to a conversation between native speakers	Global understanding, and understanding viewpoints	Informal conversation about renting a flat (approx. 1.5 minutes)	True or false	6
3	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording; selective listening	Radio programme about a museum (approx. 2.5 minutes)	Multiple choice (3 options)	6
4	Listening to video recordings (ATAO)	Understanding the main points of the video	Short about daily news (approx. 1.5 minutes)	Multiple choice questions with two alternatives; completing the statement with the correct word	6

Appendix 2A – Table B: Main characteristics of the B1 English language paper-based task set

Table B

Main characteristics of the B1 English language paper-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Global understanding and understanding one piece of specific information	Short announcements and instructions (approx. 2.5 minutes/1 recording; altogether 3 recordings)	1 true or false item for the main message per recording 1 multiple choice question (3 options) for a specific detail per recording	6
2	Listening to a conversation between native speakers	Global understanding, and understanding viewpoints	An interview about sports (approx. 2.5 minutes)	True or false	6
3	Listening to a public speech	Understanding the main points of the recording; selective listening	Radio programme about tourist guides (approx. 2 minutes)	Fill-in the gaps; completing notes with 2-3 words	6
4	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording; selective listening	Radio programme about flying vehicles (approx. 2.5 minutes)	Multiple choice (3 options)	6
5	Listening to video recordings (ATAO)	Understanding the main points of the video	Short videos about different informal topics	Multiple choice (3 options)	5

Appendix 3A – Table C: Main characteristics of the B2 English language paper-based task set

Table C

Main characteristics of the B2 English language paper-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Understanding specific information	Short announcements and instructions (approx. 3 minutes altogether)	Multiple choice (3 options)	6
2	Listening to a conversation between native speakers	Understanding viewpoints and specific information	An interview with a teacher (approx. 2.5 minutes)	True or false	6
3	Listening to a public speech	Understanding main points of the recording, selective listening	Radio programme about historic houses (approx. 2 minutes)	Fill-in the gaps; completing notes with 2-3 words	6
4	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording, selective listening	Radio programme about challenges in life (approx. 3 minutes)	Multiple choice (3 options)	6
5	Listening to video recordings (ATAO)	Understanding the main points of the recording, selective listening	Short videos about different informal topics (approx. 4 minutes altogether)	Multiple choice (3 options)	5

Appendix 4A – Table D: Main characteristics of the C1 English language paper-based task set

Table D

Main characteristics of the C1 English language paper-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Understanding specific information	Short announcements and instructions (approx. altogether 3 minutes)	Multiple choice (3 options)	6
2	Listening to a conversation between native speakers	Understanding viewpoints and specific information	A short conversation about fashion (approx. 3 minutes)	True or false	6
3	Listening to a public speech	Understanding main points of the recording, selective listening	A short presentation on organic food (approx. 2 minutes)	Fill-in the gaps; completing notes with 2-3 words	7
4	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording, selective listening	A radio programme about guide dogs (approx. 3 minutes)	Answering short questions with maximum 3 words	5
5	Listening to video recordings (ATAO)	Understanding the main points of the recording, selective listening	Short videos about daily news (approx. 4 minutes altogether)	Multiple choice (3 options)	5

Appendix 5A – Table E: Main characteristics of the A2 English language computer-based task set

Table E

Main characteristics of the A2 English language computer-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Global understanding and understanding one piece of specific information	Short airport announcements and instructions (approx. 2.5 minutes/1 recording; altogether 3 recordings)	Fill-in the gaps; completing notes with 1 word	6
2	Listening to a conversation between native speakers	Global understanding, and understanding viewpoints	Informal conversation about renting a flat (approx. 1.5 minutes)	True or false	7
3	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording; selective listening	Radio programme about a museum (approx. 2.5 minutes)	Multiple choice (3 options)	6
4	Watching video recordings (audio-visual input)	Understanding the main points of the video	Short video about daily news (approx. 1.5 minutes)	Multiple choice questions with two alternatives; completing the statement with the correct word	6

Appendix 6A – Table F: Main characteristics of the B1 English language computer-based task set

Table F

Main characteristics of the B1 English language computer-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Global understanding and understanding one piece of specific information	Short weather announcements and instructions (approx. 2.5 minutes/1 recording; altogether 3 recordings)	1 true or false item for the main message per recording 1 multiple choice question (3 options) for a specific detail per recording	6
2	Listening to a conversation between native speakers	Global understanding, and understanding viewpoints	An interview sport (approx. 2 minutes)	True or false	6
3	Listening to a public speech	Understanding the main points of the recording; selective listening	Radio programme about a tour guide (approx. 2 minutes)	Fill-in the gaps; completing notes with 2-3 words	5
4	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording; selective listening	Radio programme about dress code (approx. 4 minutes)	Multiple choice (3 options)	6
5	Watching video recordings (audio-visual input)	Understanding the main points of the video	Short videos about different informal topics	Multiple choice (3 options)	5

Appendix 7A – Table G: Main characteristics of the B2 English language computer-based task set

Table G

Main characteristics of the B2 English language computer-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Understanding specific information	Short announcements and instructions (approx. 3 minutes altogether)	Multiple choice (3 options)	6
2	Listening to a conversation between native speakers	Understanding viewpoints and specific information	A conversation about business clothes (approx. 2 minutes)	True or false	5
3	Listening to a public speech	Understanding main points of the recording, selective listening	A short presentation about ancient chariots (approx. 2 minutes)	Fill-in the gaps; completing notes with 2-3 words	6
4	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording, selective listening	Radio programme about challenges in life (approx. 1.5 minutes)	Multiple choice (3 options)	5
5	Watching video recordings (audio-visual input)	Understanding the main points of the recording, selective listening	Short videos about different informal topics (approx. 4 minutes altogether)	Multiple choice (3 options)	5

Appendix 8A – Table H: Main characteristics of the C1 English language computer-based task set

Table H

Main characteristics of the C1 English language computer-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Understanding specific information	Short announcements and instructions (approx. altogether 3 minutes)	Multiple choice (3 options)	6
2	Listening to a conversation between native speakers	Understanding viewpoints and specific information	A short conversation about talking to ex boyfriends/girlfriends (approx. 3 minutes)	True or false	6
3	Listening to a public speech	Understanding main points of the recording, selective listening	A short presentation on smart homes (approx. 3 minutes)	Fill-in the gaps; completing notes with 2-3 words	6
4	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording, selective listening	A radio programme about time management (approx. 3 minutes)	Answering short questions with maximum 3 words	8
5	Watching video recordings (audio-visual input)	Understanding the main points of the recording, selective listening	Short videos about different informal topics (approx. 4 minutes altogether)	Multiple choice (3 options)	6

Appendix 9A – Table I: Main characteristics of the A2 German language paper-based task set

Table I

Main characteristics of the A2 German language paper-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Global understanding and understanding one piece of specific information	Short airport announcements and instructions (approx. 2.5 minutes/1 recording; altogether 3 recordings)	1 true or false item for the main message per recording 1 multiple choice question (3 options) for a specific detail per recording	6
2	Listening to a conversation between native speakers	Global understanding, and understanding viewpoints	Informal conversation about renting a flat (approx. 1.5 minutes)	True or false	6
3	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording; selective listening	Radio programme about a museum (approx. 2.5 minutes)	Multiple choice (3 options)	6
4	Listening to video recordings (ATAO)	Understanding the main points of the video	Short about daily news (approx. 1.5 minutes)	Multiple choice questions with two alternatives; completing the statement with the correct word	6

Appendix 10A – Table J: Main characteristics of the B1 German language paper-based task set

Table J

Main characteristics of the B1 German language paper-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Global understanding and understanding one piece of specific information	Short announcements and instructions (approx. 2.5 minutes/1 recording; altogether 3 recordings)	1 true or false item for the main message per recording 1 multiple choice question (3 options) for a specific detail per recording	6
2	Listening to a conversation between native speakers	Global understanding, and understanding viewpoints	An interview about the ideal father (approx. 2.5 minutes)	True or false	6
3	Listening to a public speech	Understanding the main points of the recording; selective listening	Radio programme about animal apartments (approx. 2 minutes)	Fill-in the gaps; completing notes with 2-3 words	6
4	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording; selective listening	Radio programme about dinner (approx. 2.5 minutes)	Multiple choice (3 options)	6
5	Listening to video recordings (ATAO)	Understanding the main points of the video	Short videos about different informal topics	Multiple choice (3 options)	5

Appendix 11A – Table K: Main characteristics of the B2 German language paper-based task set

Table K

Main characteristics of the B2 German language paper-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Understanding specific information	Short announcements and instructions (approx. 3 minutes altogether)	Multiple choice (3 options)	6
2	Listening to a conversation between native speakers	Understanding viewpoints and specific information	An interview about the ideal father (approx. 2.5 minutes)	True or false	6
3	Listening to a public speech	Understanding main points of the recording, selective listening	Radio programme about animal apartments (approx. 2 minutes)	Fill-in the gaps; completing notes with 2-3 words	6
4	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording, selective listening	Radio programme about consuming sugar (approx. 3 minutes)	Multiple choice (3 options)	6
5	Listening to video recordings (ATAO)	Understanding the main points of the recording, selective listening	Short videos about different informal topics (approx. 4 minutes altogether)	Multiple choice (3 options)	5

Appendix 12A – Table L: Main characteristics of the C1 German language paper-based task set

Table L

Main characteristics of the C1 German language paper-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Understanding specific information	Short announcements and instructions (approx. altogether 3 minutes)	Multiple choice (3 options)	6
2	Listening to a conversation between native speakers	Understanding viewpoints and specific information	A short conversation about the life of monkeys (approx. 3 minutes)	True or false	6
3	Listening to a public speech	Understanding main points of the recording, selective listening	A short presentation on a German film (approx. 2 minutes)	Fill-in the gaps; completing notes with 2-3 words	7
4	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording, selective listening	A radio programme about online gambling games (approx. 3 minutes)	Answering short questions with maximum 3 words	5
5	Listening to video recordings (ATAO)	Understanding the main points of the recording, selective listening	Short videos about different informal topics (approx. 4 minutes altogether)	Multiple choice (3 options)	5

Appendix 13A – Table M: Main characteristics of the A2 German language computer-based task set

Table M

Main characteristics of the A2 German language computer-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Global understanding and understanding one piece of specific information	Short announcements and instructions (approx. 2.5 minutes/1 recording; altogether 3 recordings)	Fill-in the gaps; completing notes with 1 word	6
2	Listening to a conversation between native speakers	Global understanding, and understanding viewpoints	Informal conversation in a café (approx. 1.5 minutes)	True or false	7
3	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording; selective listening	Radio programme about searching for a flat (approx. 2.5 minutes)	Multiple choice (3 options)	6
4	Watching video recordings (audio-visual input)	Understanding the main points of the video	Short video about mobile phones (approx. 1.5 minutes)	Multiple choice questions with two alternatives; completing the statement with the correct word	6

Appendix 14A – Table N: Main characteristics of the B1 German language computer-based task set

Table N

Main characteristics of the B1 German language computer-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Global understanding and understanding one piece of specific information	Short announcements and instructions (approx. 2.5 minutes/1 recording; altogether 3 recordings)	1 true or false item for the main message per recording 1 multiple choice question (3 options) for a specific detail per recording	6
2	Listening to a conversation between native speakers	Global understanding, and understanding viewpoints	An interview with an animal trainer (approx. 2 minutes)	True or false	6
3	Listening to a public speech	Understanding the main points of the recording; selective listening	Radio programme about travelling tips (approx. 2 minutes)	Fill-in the gaps; completing notes with 2-3 words	5
4	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording; selective listening	Radio programme about riding bicycles (approx. 4 minutes)	Multiple choice (3 options)	6
5	Watching video recordings (audio-visual input)	Understanding the main points of the video	Short videos about different informal topics	Multiple choice (3 options)	5

Appendix 15A – Table O: Main characteristics of the B2 German language computer-based task set

Table O

Main characteristics of the B2 German language computer-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Understanding specific information	Short announcements and instructions (approx. 3 minutes altogether)	Multiple choice (3 options)	6
2	Listening to a conversation between native speakers	Understanding viewpoints and specific information	A conversation about food swapping (approx. 2 minutes)	True or false	5
3	Listening to a public speech	Understanding main points of the recording, selective listening	A short presentation about migraine (approx. 2 minutes)	Fill-in the gaps; completing notes with 2-3 words	6
4	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording, selective listening	Radio programme about the connection between intelligence and behaviour (approx. 1.5 minutes)	Multiple choice (3 options)	5
5	Watching video recordings (audio-visual input)	Understanding the main points of the recording, selective listening	Short videos about different informal topics (approx. 4 minutes altogether)	Multiple choice (3 options)	5

Appendix 16A – Table P: Main characteristics of the C1 German language computer-based task set

Table P

Main characteristics of the C1 German language computer-based task set

Task	Listening activities	Comprehension strategies	Text type and length	Task type	Number of Items
1	Listening to public announcements (information, instructions, warnings, etc.)	Understanding specific information	Short announcements and instructions (approx. altogether 3 minutes)	Multiple choice (3 options)	6
2	Listening to a conversation between native speakers	Understanding viewpoints and specific information	A short conversation about being an adult (approx. 3 minutes)	True or false	6
3	Listening to a public speech	Understanding main points of the recording, selective listening	A short presentation on learning (approx. 3 minutes)	Fill-in the gaps; completing notes with 2-3 words	6
4	Listening to media (radio, TV, recordings, cinema)	Understanding the main points of the recording, selective listening	A radio programme about measuring blood pressure (approx. 3 minutes)	Answering short questions with maximum 3 words	8
5	Watching video recordings (audio-visual input)	Understanding the main points of the recording, selective listening	Short videos about different informal topics (approx. 4 minutes altogether)	Multiple choice (3 options)	6

Appendix 1B – Think-aloud tasks – the original Hungarian version and the English translation

Demonstration think-aloud task

Alkoss az alábbi összekevert szavakból az összes szó felhasználásával egy értelmes és nyelvtanilag helyes angol mondatot. Közben amennyire lehet, kérlek, verbalizáld minden gondolatodat úgy, mintha megpróbálnál végigvezetni engem a megoldási folyamaton. Gondolataidat angolul, vagy magyarul is verbalizálhatod. Kérlek, használd azt a nyelvet, amelyiken a gondolat megfogalmazódik a fejedben.

[Create a meaningful and grammatically correct English sentence from the words below. Please, use all the words, and while solving the task, verbalise every single thought that emerges in your mind. You can verbalise your thoughts both in English and Hungarian. Please use the language you are thinking in.]

Szavak [Words]: Even a he he's lot makes mistakes of still thoroughly though trained

Mintamegoldás [Sample solution]: Even though he's thoroughly trained, he still makes a lot of mistakes.

Practice think-aloud task #1

Alkoss az alábbi összekevert szavakból az összes szó felhasználásával egy értelmes és nyelvtanilag helyes angol mondatot. Közben amennyire lehet, kérlek, verbalizáld minden gondolatodat úgy, mintha megpróbálnál végigvezetni engem a megoldási folyamaton. Gondolataidat angolul, vagy magyarul is verbalizálhatod. Kérlek, használd azt a nyelvet, amelyiken a gondolat megfogalmazódik a fejedben.

[Create a meaningful and grammatically correct English sentence from the words below. Please, use all the words, and while solving the task, verbalise every single thought that emerges in your mind. You can verbalise your thoughts both in English and Hungarian. Please use the language you are thinking in.]

Szavak [Words]: You service in take because a starting it's for your should car to make weird noises

Mintamegoldás [Sample solution]: You should take your car in for a service because it's starting to make weird noises.

Practice think-aloud task #2

Alkoss az alábbi összekevert szavakból az összes szó felhasználásával egy értelmes és nyelvtanilag helyes angol mondatot. Közben amennyire lehet, kérlek, verbalizáld minden gondolatodat úgy, mintha megpróbálnál végigvezetni engem a megoldási folyamaton. Gondolataidat angolul, vagy magyarul is verbalizálhatod. Kérlek, használd azt a nyelvet, amelyiken a gondolat megfogalmazódik a fejedben.

[Create a meaningful and grammatically correct English sentence from the words below. Please, use all the words, and while solving the task, verbalise every single thought that emerges in your mind. You can verbalise your thoughts both in English and Hungarian. Please use the language you are thinking in.]

Szavak [Words]: I I coupon to come case to back in save have this the store tomorrow

Mintamegoldás [Sample solution]: I have to save this coupon in case I come back to the store tomorrow.

Appendix 2B – Semi-structured interview – the original Hungarian version

Semi-structured interview in Hungarian for the participants piloting the and audio-only and audio-visual tasks

Félig strukturált interjúterv

Kedves Vizsgázó! A következő interjúban a hallott szöveg értése és az audiovizuális szöveg értése feladatokkal kapcsolatos tapasztalataidra vagyok kíváncsi. Mivel nincsenek jó vagy rossz válaszok, kérlek, próbálj meg minél őszintébben válaszolni a feltett kérdésekre. A felvett adatok a doktori disszertációs kutatásom részét képezik majd, melyeket anonim módon – név nélkül – kezelek. Tudnod kell, hogy a válaszadás önkéntes és bármikor meggondolhatod magad akár az interjú közben is. Nagyon szépen köszönöm, hogy válaszaiddal hozzájárulsz kutatásom adatgyűjtési részéhez!

Kapcsolatteremtő kérdések:

1. Hány éves vagy?
2. Mi a foglalkozásod?
3. Mióta tanulod az angolt/németet?
4. Heti hány órában tanulod az angolt/németet a nyelviskolában/iskolában?
5. Ezen a nyelviskolai kurzuson kívül, tanulod még valahol szervezett intézményi keretek között (pl. céges tanfolyamon, másik nyelviskolában, magántanárral stb.) az angolt/németet?
6. Milyen más idegennyelveket tanulsz/tanultál az angolon/németen kívül?
7. Milyen szerepet játszik az angol/német nyelv a jövőbeni terveidben?

A hallott szöveg értése feladatokra vonatkozó kérdések:

1. Mennyire találtad nehéznek a hallott szövegértés feladatokat? Mit találtál nehéznek bennük?
2. Voltak-e olyan szavak vagy kifejezések a hanganyagban vagy a feladatokban, amiknek a meg nem értése akadályozott a feladat megoldásában? Melyik feladatoknál volt így, s ez hogyan befolyásolta a válaszadásodat?
3. Elegendőnek találtad a feladatokra megadott időt? Mennyi idő lett volna számodra ideális?
4. Mit gondolsz a feladatok számáról? Kevésnek vagy soknak találod a feladatok mennyiségét?
5. Mennyire találtad érdekesnek a feladatokat? Mit találtál bennük érdekesnek?
6. Volt olyan kérdés a feladatokban, amit a szöveg meghallgatása nélkül is meg tudtál volna válaszolni?
7. Találkoztál-e bármilyen zavaró tényezővel a feladatok megoldása során? Amennyiben volt, zavart-e a háttérzaj a hangfelvételen? Zavart-e a beszélő hangszíne/hanghordozása a hangfelvételen? Zavart-e a beszélő akcentusa a hangfelvételen?
8. Van esetleg bármilyen egyéb észrevételed a hallott szöveg értés feladatokkal kapcsolatban?

Az audiovizuális szöveg értése feladatokra vonatkozó kérdések:

1. Mennyire találtad nehéznek az audiovizuális szövegértés feladatokat? Mit találtál nehéznek bennük?
2. Voltak-e olyan szavak vagy kifejezések a videókban vagy a feladatokban, amiknek a meg nem értése akadályozott a feladat megoldásában? Ez hogyan befolyásolta a válaszadásodat?
3. Elegendőnek találtad a feladatokra megadott időt? Mennyi idő lett volna számodra ideális?
4. Mit gondolsz a feladat számáról? Hasznos lett volna, több videós feladat?
5. Hasznosnak találtad-e, hogy a feladatokhoz nem csak audiofelvétel, hanem videó is volt? Milyen szempontból találtad hasznosnak/feleslegesnek?
6. Melyik feladatokat találtad könnyebbnek: a hallott szövegértés feladatokat vagy az audiovizuális szövegértés feladatokat? Kérlek, indokold meg a válaszod.
7. Mit csináltál másképp az audiovizuális feladatok megoldása során a hagyományos hallott szöveg értése feladatokhoz képest?
8. Ha lett volna rá lehetőséged, megállítottad-e volna a videókat? Miért (nem)?
9. Ha lett volna rá lehetőséged, visszatekertél-e volna valahol a videókban? Hova és miért (nem)?
10. Mennyire találtad érdekesnek a feladatokat? Mit találtál bennük érdekesnek?
11. Volt olyan kérdés a feladatokban, amit a videók megnézése nélkül is meg tudtál volna válaszolni?
12. Találkoztál bármilyen zavaró tényezővel a feladatok megoldása során? Zavart-e a beszélő hangszíne/hanghordozása a videofelvételen? Zavart-e a beszélő akcentusa a videofelvételen?
13. Van esetleg bármilyen egyéb észrevételed az audiovizuális szöveg értés feladatokkal kapcsolatban?

Végül pedig fontos még megkérdeznem, hogy hozzájárulsz-e, hogy válaszaidat a kutatás céljára felhasználjam.

Appendix 3B – Semi-structured interview – the English translation

Semi-structured interview in English for the participants piloting the audio-only and audio-visual tasks

Semi-structured interview schedule

Dear Examinee,
In the following interview I would like to ask you about your experience in connection with the listening and audio-visual tasks. There are no right or wrong answers. Therefore, I would like you to be as honest as possible regarding your answers to the questions. The data you provide is going to be part of my doctoral dissertation. Your data is handled in an anonymous way. Taking part in this interview is voluntary and you can decline to answer any questions during the interview. Thank you very much for your help.

Building a rapport:

1. How old are you?
2. What is your occupation?
3. How long have you been learning English/German?
4. How many English/German classes do you have in a week at the (language) school?
5. Besides having this language course, do you learn English/German in a formal institutionalised way (e.g., language course at the office, courses at another language school, private tutoring etc.)?
6. What other foreign languages have you learnt/ do you learn besides English/German?
7. What are your future plans in connection with the English/German language?

Questions regarding the listening tasks:

1. How difficult have you found the listening tasks? What have you found difficult in them?
2. Were there any words or expressions in the recordings of the tasks which made it difficult to answer a question? Which task was this? How did these words/expressions influence your answer?
3. Was the time provided to solve the tasks enough? How much time would you find appropriate to have?
4. What is your opinion about the number of tasks? Have you found them to be few or many?
5. To what extent do you think the tasks were interesting? What have you found interesting in them?
6. Were there any items which you could solve even without listening to the recording?
7. Were there anything disturbing for you while solving the tasks? Was there any background noise in the recording which you found to be disturbing? Was the tone of the speakers disturbing for you? Was the accent of the speakers disturbing for you?
8. Is there anything else you think is important to discuss in connection with the listening tasks?

Questions regarding the audio-visual tasks:

1. How difficult have you found the audio-visual tasks? What have you found difficult in them?
2. Were there any words or expressions in the video of the tasks which made it difficult to answer a question? How did these words/expressions influence your answer?
3. Was the time provided to solve the task enough? How much time would you find appropriate to have?
4. What is your opinion about the number of tasks? Do you think more audio-visual tasks would be better?
5. Have you found it useful to have video recordings in the task? From what point of view was it useful?
6. Which tasks were easier for you to answer: the listening tasks or the audio-visual tasks? Please justify your answer.
7. What did you do differently in solving the audio-visual tasks comparing it to solving the listening tasks?
8. Would you have stopped the video recording if there had been the opportunity? Why (not)?
9. Would you have rewound the video recording if there had been the opportunity? To which point of the recording would you have rewound it and why? Why not?
10. To what extent do you think the task was interesting? What have you found interesting in it?
11. Were there any items which you could solve even without watching the video?
12. Were there anything disturbing for you while solving the tasks? Was the tone of the speakers disturbing for you in the video? Was the accent of the speakers disturbing for you in the video?
13. Is there anything else you think is important to discuss in connection with the audio-visual task?

Finally, I would like to ask whether you agree to use your answers in my research study.

Appendix 1C – Questionnaire in Hungarian about the Paper-Based Tests

Kérdőív – Preteszt Vizsgázók Számára

A nevem Kövér Ármin, az Eötvös Loránd Tudományegyetem, Nyelvpedagógiai Doktori programjának hallgatója vagyok. A doktori disszertációs kutatásom adatgyűjtését végzem. Az alábbi kérdőív kitöltésében szeretném a segítségedet. A kérdőívben a hallott szöveg értése és az audiovizuális szöveg értése feladatokhoz fűződő tapasztalataidra vagyok kíváncsi. Minden esetben a Te saját véleményedre vagyok kíváncsi. Kérlek, őszintén válaszolj! Próbálj meg minden kérdésre válaszolni, még ha kicsit bizonytalan is vagy! A kérdőív kitöltése önkéntes és anonim módon történik, és körülbelül 10 percet vesz igénybe. Válaszaidat bizalmasan kezelem. Segítségedet nagyon szépen köszönöm!

A kérdőívnek ebben a részében arra szeretnélek kérni, hogy válaszolj az alábbi állításokra egy-egy szám (1–5) bekarikázásának segítségével.

Például, ha „határozottan nem értesz egyet" azzal az állítással, hogy *a matekdolgozat feladata könnyű volt*, „részben egyetértesz, részben nem" azzal az állítással, hogy *sok képletet kellett ismerni a feladat megoldásához*, illetve „határozottan egyetértesz" azzal az állítással, hogy *hasznos volt, hogy lehetett számológépet használni a feladat megoldásához*, akkor a válaszaidat így jelöld:

	Határozottan nem értek egyet	**Nem igazán értek egyet**	**Részben egyetértek, részben nem**	**Nagyjából egyetértek**	**Határozottan egyetértek**
01. A matekdolgozat feladata könnyű volt.	**1**	**2**	**3**	**4**	**5**
02. Sok képletet kellet ismerni a feladat megoldásához.	**1**	**2**	**3**	**4**	**5**
03. Hasznos volt, hogy lehetett számológépet használni.	**1**	**2**	**3**	**4**	**5**

Kérlek, **csak 1db számot** jelölj meg minden sorban!

	Határozottan nem értek egyet	**Nem igazán értek egyet**	**Részben egyetértek, részben nem**	**Nagyjából egyetértek**	**Határozottan egyetértek**
1. A feltett kérdések száma elegendő volt a szövegek hosszához viszonyítva.	**1**	**2**	**3**	**4**	**5**
2. Kellemetlen volt a beszélők hangszíne.	**1**	**2**	**3**	**4**	**5**
3. A szövegek gondolatmenetei jól követhetők voltak.	**1**	**2**	**3**	**4**	**5**
4. A feladatok összességében nehezek voltak.	**1**	**2**	**3**	**4**	**5**

5. A felvételeken hallható háttérzaj zavaró volt.	**1**	**2**	**3**	**4**	**5**
6. A feladatokra szánt idő elegendő volt.	**1**	**2**	**3**	**4**	**5**
7. A beszélők akcentusa érthetetlen volt.	**1**	**2**	**3**	**4**	**5**
8. A szövegekből sok mindenre kellett egyszerre emlékezni ahhoz, hogy meg lehessen válaszolni a kérdéseket.	**1**	**2**	**3**	**4**	**5**
9. A lejátszások száma elegendő volt.	**1**	**2**	**3**	**4**	**5**
10. A szövegek gyorsasága megfelelő volt.	**1**	**2**	**3**	**4**	**5**
11. A felvételeken hallható zenék zavarók voltak.	**1**	**2**	**3**	**4**	**5**
12. A feladatban voltak olyan kérdések, amikre a választ pusztán tippeltem.	**1**	**2**	**3**	**4**	**5**
13. Jobb lett volna, ha vannak videók is a szövegekhez.	**1**	**2**	**3**	**4**	**5**
14. A szövegek helyenként tartalmaztak olyan szókincset, amit nem ismertem.	**1**	**2**	**3**	**4**	**5**
15. Egy-egy videó segíthette volna a hangzó szövegek jobb megértését.	**1**	**2**	**3**	**4**	**5**
16. A feladatok szövegeiben volt számomra ismeretlen szókincs.	**1**	**2**	**3**	**4**	**5**
17. Egy-egy videó segíthette volna a hangzó szövegek gondolatmenetének követését.	**1**	**2**	**3**	**4**	**5**
18. Egy-egy videó segíthette volna a feladatok megválaszolását.	**1**	**2**	**3**	**4**	**5**

Nemed (Jelöld „X”-szel.):

Fiú	
Lány	

Az általad írt teszt szintje. (Jelöld „X”-szel.)

A2	
B1	
B2	
C1	

Az általad írt teszt verziója (Jelöld „X”-szel.):

Papíralapú	
Digitális	

Appendix 2C – Questionnaire about the Paper-Based Tests – English translation

Questionnaire – For Pre-test Examinees

My name is Ármin Kövér and I am a PhD student at Eötvös Loránd University. I am collecting data for my PhD dissertation and I would like to ask your help by completing the following questionnaire. In the questionnaire, you have to answer questions related to the listening and audio-visual tasks based on your experience. I am interested in your opinion. Please, answer the questions as honestly as possible. Please, answer all the questions even if you are not sure about your answers. Filling in the questionnaire is happening in a voluntary and anonymous way. Filling in the questionnaire takes approximately 10 minutes. Your answers are handled confidentially. Thank you very much for your help.

In this part of the questionnaire, I would like you to answer the following questions by circling a number (1–5).

For example, if you "strongly disagree" with the statement that *the task of the math test was easy*; you "partly agree, partly disagree" with the statement that *it was necessary to know many formulae to solve the task*; and you "strongly agree" with the statement that *it was useful to use a calculator*, please indicate your answers in the following way:

	Strongly disagree	**Disagree**	**Partly agree, partly disagree**	**Agree**	**Strongly agree**
01. The task of the math test was easy.	**1**	**2**	**3**	**4**	**5**
02. It was necessary to know many formulae to solve the task.	**1**	**2**	**3**	**4**	**5**
03. It was useful to use a calculator.	**1**	**2**	**3**	**4**	**5**

Please, circle **only one number** in every row.

	Strongly disagree	**Disagree**	**Partly agree, partly disagree**	**Agree**	**Strongly agree**
1. The number of questions was enough compared to the length of the recordings.	**1**	**2**	**3**	**4**	**5**
2. The tone of the speakers was disturbing.	**1**	**2**	**3**	**4**	**5**
3. I could follow the line of thoughts of the recordings.	**1**	**2**	**3**	**4**	**5**
4. Overall, the tasks were difficult.	**1**	**2**	**3**	**4**	**5**

5. The background noise on the recordings was disturbing.	**1**	**2**	**3**	**4**	**5**
6. The time provided to solve the tasks was enough.	**1**	**2**	**3**	**4**	**5**
7. The accent of the speakers was incomprehensible.	**1**	**2**	**3**	**4**	**5**
8. I had to remember several information at the same time to answer the questions.	**1**	**2**	**3**	**4**	**5**
9. The number of playing the recordings was enough.	**1**	**2**	**3**	**4**	**5**
10. The pace of the recordings was appropriate.	**1**	**2**	**3**	**4**	**5**
11. The music on the recordings was disturbing.	**1**	**2**	**3**	**4**	**5**
12. I guessed some answers to some of the questions.	**1**	**2**	**3**	**4**	**5**
13. It could have been better if there had been a video to some of the recordings.	**1**	**2**	**3**	**4**	**5**
14. Some of the recordings contained words that I did not know.	**1**	**2**	**3**	**4**	**5**
15. A video could have helped the better understanding of some of the audio recordings.	**1**	**2**	**3**	**4**	**5**
16. There were unfamiliar words in the recordings.	**1**	**2**	**3**	**4**	**5**
17. A video could have helped following the lines of thoughts in some of the recordings.	**1**	**2**	**3**	**4**	**5**
18. A video could have helped answer the questions in some of the tasks.	**1**	**2**	**3**	**4**	**5**

Your gender (mark with an "X"):

Male	
Female	

The level of the test (mark with an "X"):

A2	
B1	
B2	
C1	

The version of the test (mark with an "X"):

Paper	
Digital	

Appendix 1D – Questionnaire in Hungarian about the Computer-Based Tests

Kérdőív – Preteszt Vizsgázók Számára

A nevem Kövér Ármin, az Eötvös Loránd Tudományegyetem, Nyelvpedagógiai Doktori programjának hallgatója vagyok. A doktori disszertációs kutatásom adatgyűjtését végzem. Az alábbi kérdőív kitöltésében szeretném a segítségedet. A kérdőívben a hallott szöveg értése és az audiovizuális szöveg értése feladatokhoz fűződő tapasztalataidra vagyok kíváncsi. Minden esetben a Te saját véleményedre vagyok kíváncsi. Kérlek, őszintén válaszolj! Próbálj meg minden kérdésre válaszolni, még ha kicsit bizonytalan is vagy! A kérdőív kitöltése önkéntes és anonim módon történik, és körülbelül 10 percet vesz igénybe. Válaszaidat bizalmasan kezelem. Segítségedet nagyon szépen köszönöm!

I. A kérdőívnek ebben a részében arra szeretnénk kérni, hogy válaszolj az alábbi állításokra egy-egy szám (1–5) bekarikázásának segítségével.

Például, ha „határozottan nem értesz egyet" azzal az állítással, hogy *a matekdolgozat feladata könnyű volt*, „részben egyetértesz, részben nem" azzal az állítással, hogy *sok képletet kellett ismerni a feladat megoldásához*, illetve „határozottan egyetértesz" azzal az állítással, hogy *hasznos volt, hogy lehetett számológépet használni a feladat megoldásához*, akkor a válaszaidat így jelöld:

	Határozottan nem értek egyet	**Nem igazán értek egyet**	**Részben egyetértek, részben nem**	**Nagyjából egyetértek**	**Határozottan egyetértek**
01. A matekdolgozat feladata könnyű volt.	**1**	**2**	**3**	**4**	**5**
02. Sok képletet kellet ismerni a feladat megoldásához.	**1**	**2**	**3**	**4**	**5**
03. Hasznos volt, hogy lehetett számológépet használni.	**1**	**2**	**3**	**4**	**5**

Kérlek, **csak 1db számot** jelölj meg minden sorban!

	Határozottan nem értek egyet	**Nem igazán értek egyet**	**Részben egyetértek, részben nem**	**Nagyjából egyetértek**	**Határozottan egyetértek**
1. A feltett kérdések száma elegendő volt a szövegek hosszához viszonyítva.	**1**	**2**	**3**	**4**	**5**
2. Kellemetlen volt a beszélők hangszíne.	**1**	**2**	**3**	**4**	**5**
3. A szövegek gondolatmenetei jól követhetők voltak.	**1**	**2**	**3**	**4**	**5**
4. A feladatok összességében nehezek voltak.	**1**	**2**	**3**	**4**	**5**
5. A felvételeken hallható háttérzaj zavaró volt.	**1**	**2**	**3**	**4**	**5**

6. A feladatokra szánt idő elegendő volt.	**1**	**2**	**3**	**4**	**5**
7. A beszélők akcentusa érthetetlen volt.	**1**	**2**	**3**	**4**	**5**
8. A szövegekből sok mindenre kellett egyszerre emlékezni ahhoz, hogy meg lehessen válaszolni a kérdéseket.	**1**	**2**	**3**	**4**	**5**
9. A lejátszások száma elegendő volt.	**1**	**2**	**3**	**4**	**5**
10. A szövegek gyorsasága megfelelő volt.	**1**	**2**	**3**	**4**	**5**
11. A felvételeken hallható zenék zavarók voltak.	**1**	**2**	**3**	**4**	**5**
12. A feladatban voltak olyan kérdések, amikre a választ pusztán tippeltem.	**1**	**2**	**3**	**4**	**5**
13. A videó segítette az utolsó szöveg jobb megértését.	**1**	**2**	**3**	**4**	**5**
14. A szövegek helyenként tartalmaztak olyan szókincset, amit nem ismertem.	**1**	**2**	**3**	**4**	**5**
15. A videó megkönnyítette az utolsó szöveg gondolatmenetének követését.	**1**	**2**	**3**	**4**	**5**
16. A feladatok szövegeiben volt számomra ismeretlen szókincs.	**1**	**2**	**3**	**4**	**5**
17. A videóban szereplő képi információk segítették az utolsó feladat megválaszolását.	**1**	**2**	**3**	**4**	**5**
18. A videót hasznosnak találtam az utolsó feladat megoldásához.	**1**	**2**	**3**	**4**	**5**

Nemed (Jelöld „X”-szel.):

Fiú	
Lány	

Az általad írt teszt szintje. (Jelöld „X”-szel.):

A2	
B1	
B2	
C1	

Az általad írt teszt verziója (Jelöld „X”-szel.):

Papíralapú	
Digitális	

Appendix 2D – Questionnaire about the Computer-Based Tests – English translation

Questionnaire – For Pre-test Examinees

My name is Ármin Kövér and I am a PhD student at Eötvös Loránd University. I am collecting data for my PhD dissertation and I would like to ask your help by completing the following questionnaire. In the questionnaire, you have to answer questions related to the listening and audio-visual tasks based on your experience. I am interested in your opinion. Please, answer the questions as honestly as possible. Please, answer all the questions even if you are not sure about your answers. Filling in the questionnaire is happening in a voluntary and anonymous way. Filling in the questionnaire takes approximately 10 minutes. Your answers are handled confidentially. Thank you very much for your help.

In this part of the questionnaire, I would like you to answer the following questions by circling a number (1–5).

For example, if you "strongly disagree" with the statement that *the task of the math test was easy*; you "partly agree, partly disagree" with the statement that *it was necessary to know many formulae to solve the task*; and you "strongly agree" with the statement that *it was useful to use a calculator*, please indicate your answers in the following way:

	Strongly disagree	**Disagree**	**Partly agree, partly disagree**	**Agree**	**Strongly agree**
01. The task of the math test was easy.	**1**	**2**	**3**	**4**	**5**
02. It was necessary to know many formulae to solve the task.	**1**	**2**	**3**	**4**	**5**
03. It was useful to use a calculator.	**1**	**2**	**3**	**4**	**5**

Please, circle **only one number** in every row.

	Strongly disagree	**Disagree**	**Partly agree, partly disagree**	**Agree**	**Strongly agree**
1. The number of questions was enough compared to the length of the recordings.	**1**	**2**	**3**	**4**	**5**
2. The tone of the speakers was disturbing.	**1**	**2**	**3**	**4**	**5**
3. I could follow the line of thoughts of the recordings.	**1**	**2**	**3**	**4**	**5**
4. Overall, the tasks were difficult.	**1**	**2**	**3**	**4**	**5**

5. The background noise on the recordings was disturbing.	**1**	**2**	**3**	**4**	**5**
6. The time provided to solve the tasks was enough.	**1**	**2**	**3**	**4**	**5**
7. The accent of the speakers was incomprehensible.	**1**	**2**	**3**	**4**	**5**
8. I had to remember several information at the same time to answer the questions.	**1**	**2**	**3**	**4**	**5**
9. The number of playing the recordings was enough.	**1**	**2**	**3**	**4**	**5**
10. The pace of the recordings was appropriate.	**1**	**2**	**3**	**4**	**5**
11. The music on the recordings was disturbing.	**1**	**2**	**3**	**4**	**5**
12. I guessed some answers to some of the questions.	**1**	**2**	**3**	**4**	**5**
13. The video helped the better understanding of the last recording.	**1**	**2**	**3**	**4**	**5**
14. Some of the recordings contained words that I did not know.	**1**	**2**	**3**	**4**	**5**
15. The video made it easier to follow the lines of thoughts in the last recording.	**1**	**2**	**3**	**4**	**5**
16. There were unfamiliar words in the recordings.	**1**	**2**	**3**	**4**	**5**
17. The visual information in the video helped to better answer the questions of the last task.	**1**	**2**	**3**	**4**	**5**
18. The video was useful in answering the questions of the last task.	**1**	**2**	**3**	**4**	**5**

Your gender (mark with an "X"):

Male	
Female	

The level of the test (mark with an "X"):

A2	
B1	
B2	
C1	

The version of the test (mark with an "X"):

Paper	
Digital	

Appendix 1E – Consent form in Hungarian

BELEEGYEZŐ NYILATKOZAT
EÖTVÖS LORÁND TUDOMÁNYEGYETEM, BUDAPEST

A tanulmány címe (munkacím):	**A nyelvhasználók hallott szövegértési és audiovizuális szövegértési teljesítményeinek vizsgálata**
A kutató neve:	**Kövér Ármin**
A program neve:	**Nyelvpedagógia Doktori Program**

Bevezető

- A jelen kutatásban a doktori disszertációs munkámhoz gyűjtök adatokat.
- Kérlek, olvasd el figyelmesen a jelen beleegyező nyilatkozatot és tedd fel esetleges kérdéseidet mielőtt beleegyeznél a kutatásban való részvételbe.

A kutatás célja

- A kutatás célja, hogy megvizsgálja a nyelvhasználók hallott szövegértési és audiovizuális szövegértési teljesítményeit különböző feladatok segítségével.
- A jelen kutatás keretein belül gyűjtött adatok a kutató doktori disszertációs tanulmányának alapjául szolgálnak majd, és a továbbiakban részét képezhetik egyéb általa publikált vagy előadott tudományos munkáknak is.

Az adatfelvétel folyamata

- Amennyiben a résztvevő beleegyezik a kutatásban való részvételbe, számos hallott szövegértési és audiovizuális szövegértési feladatot kell megoldania. A feladatok megoldása után a résztvevőnek kérdésekre kell válaszolnia a feladatok megoldásának folyamatával kapcsolatban. Ez hangfelvétel formájában rögzítésre kerül.

Az adatok kezelése

- A kutatásban való részvétel anonim és önkéntes módon történik. A teljes anonimitás biztosításának érdekében, az adatok elemzése és publikálása során a részvevők álneveket kapnak. Az adatok bármilyen formájú publikálása esetén a résztvevőkkel kapcsolatban semmilyen olyan személyes adat nem kerül publikálásra, amely személyazonosságukat felismerhetővé tehetné.
- A kutatással kapcsolatos minden dokumentum, a jelen beleegyező nyilatkozatot is beleértve, szigorúan titkos. A kutatással kapcsolatos minden fizikai dokumentum egy lezárt szekrényben kerül tárolásra, az elektronikus fájlok biztonságáról pedig titkosított és jelszóval védett mappák gondoskodnak. A kutatón kívül más sem a fizikai, sem az elektronikus dokumentumokhoz nem fog hozzáféréssel rendelkezni. Öt évvel a doktori cím megszerzése után, a kutatáshoz kapcsolódó minden adat és dokumentum megsemmisítésre kerül.

Részvételtől való visszalépés

- A résztvevőnek jogában áll a kutatásban való részvételtől bármikor elállni. Az adatgyűjtési folyamat bármely pontján a résztvevőnek jogában áll bármelyik kérdés megválaszolását megtagadni vagy akár a teljes részvételtől visszalépni.

Beleegyezés

- A résztvevő alábbi aláírásával kijelenti, hogy a tanulmányban önkéntes részvételt vállal. Továbbá aláírásával tanúsítja, hogy a jelen dokumentumban foglaltakat elolvasta, megértette és elfogadja. A résztvevő az aláírt és dátummal ellátott dokumentumból egy darab másolatot kap.

A résztvevő neve (NYOMTATOTT BETŰVEL): _____

A résztvevő aláírása: _____ **Dátum:** _____

A kutató aláírása: _____ **Dátum:** _____

Appendix 2E – Consent form – English translation

CONSENT TO PARTICIPATE IN A RESEARCH STUDY
EÖTVÖS LORÁND UNIVERSITY, BUDAPEST, HUNGARY

Title of the Study (working title):	**Investigating Language Users' Performance in Listening Comprehension and Audio-visual Comprehension**
Researcher:	**Ármin Kövér**
Programme:	**PhD Programme in Language Pedagogy**

Introduction

- You are being asked to participate in a research study conducted for my PhD dissertation.
- I ask that you read this form and ask any questions that you may have before agreeing to participate in the study.

Purpose of Study

- The purpose of the study is to gather insights about the participants' opinions regarding audio-visual and listening comprehension tasks.
- Ultimately, the data obtained from this research project will be published as part of my PhD dissertation, and later it might also be published or presented as part of an academic article.

Description of the Study Procedures

- If you agree to participate in this study, you will be asked to solve a series of listening and audio-visual tasks. Having solved the tasks, you will be asked to answer a series of questions related to the tasks. This will be audio recorded.

Confidentiality

- This study is anonymous. To maintain the anonymity, when the data from this study is published in any format, the names of the participants will be changed. I will not include any information in any report I may publish that would make it possible to identify any of the participants.
- The records of this study will be kept strictly confidential. Research records will be kept in a locked file, and all electronic information will be coded and secured using a password protected file. Nobody besides me will have access to the collected data. Five years after obtaining my doctorate, all the physical and electronic documents related to this study will be permanently destroyed.

Right to Refuse or Withdraw

- The decision to participate in this study is entirely up to you. You may refuse to take part in the study *at any time* without any consequences. You have the right not to answer any single question, as well as to withdraw completely from the interview at any point during the process.

Consent

- Your signature below indicates that you have decided to volunteer as a research participant for this study, and that you have read and understood the information provided above. You will be given one signed and dated copy of this form to keep.

Participant's Name (with CAPITAL LETTERS): _____

Participant's Signature: _____ **Date:** _____

Researcher's Signature: _____ **Date:** _____