THESIS BOOKLET

ÁRMIN KÖVÉR

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

Eötvös Loránd University
Faculty of Education and Psychology
Doctoral School of Education
PhD Programme in Language Pedagogy

Supervisor: Gergely Dávid, PhD

2020

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

Table of Contents

1 Introduction

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, Douglas, Jamieson, Nissan, and Turner (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 is to analyse whether including audio-visual tasks in the listening comprehension component of language examinations is necessary and desirable.

2 Theoretical background

2.1 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)

Another definition of listening comprehension which is often considered is the CEFR's definition (Council of Europe, 2001). 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 these definitions, speech comprehension requires the listener to decode utterances. During this decoding process, the acoustic characteristics of sounds (e.g., length and loudness) help the listener to decode the different speech signals from the stream of sounds (Marslen-Wilson & Tyler, 1980). Additionally, understanding speech signals requires some time to be processed (Brazil, 1983; Chafe, 1980, 1982; Kreckel 1981), and identifying phonemic units (Chomsky & Halle, 1968). 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).

Regarding testing listening comprehension, a key issue which should be considered 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). 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.

2.2 Audiovisual comprehension

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.

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 et al. (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 methods

The aim of the dissertation is to analyse whether including audio-visual tasks into 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?

3.1 Data collection

The proposed issue was investigated 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 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.

In the first data collection phase, 16 sets of listening comprehension tasks were developed (i.e., 8 English and 8 German) for four different language proficiency levels (i.e., A2, B1, B2, C1) by a language examination development team assembled by the Hungarian language school responsible for the language examination development project. 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, Grotjahn, & Tesch, 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.

3.2 Participants

As the questionnaire development did not form part of the language examination development project, the participants of the first and second data collection phase were recruited by the author of this dissertation independently from the project. During the first phase, fifteen participants (i.e., 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. 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.

During 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. They 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. 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 the participants of the first and the second data collection phase were not taken into consideration when answering the research questions.

The third data collection phase had altogether 140 participants (i.e., 60 males and 80 females). 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 1 and Table 2.

Table 1

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 2

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

3.3 Instruments

The data collection instruments used to answer the research questions can be divided into two categories: the paper-based and computer-based sets of tasks, and the respective questionnaires. For the paper-based and computer-based tests used for data collection, altogether 8 audio-visual tasks, 8 ATAO tasks, and 60 audio-only tasks were developed. These tasks were organised into 8 sets of tasks intended for a paper-based test (i.e., 4 English and 4 German), and 8 sets of tasks intended for a computer-based test (i.e., 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. 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 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 an 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 type of data collection instrument was the questionnaire. Two versions of the questionnaire were developed during the first two phases of the data collection: a paper-based test questionnaire and a computer-based test questionnaire. 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. The final version of the questionnaire investigated four constructs: *disturbing features*, *structure of the test*, *perceived difficulty*, and *necessity of the video*. 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.

3.4 Data analysis

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. 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. The data analysis followed the classical test theory approach. This approach was chosen

over the modern test theory approaches (e.g., item response theory) because the present sample is too fragmented and too small for reliable IRT analyses.

4 Results and discussion

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 (Table 3); the A2, B1, and B2 English computer-based tests (Table 4); and the B1 and C1 German computer-based tests (Table 5).

Table 3

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.

Table 4

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.

Table 5

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.

Furthermore, the reliability of the B2 English paper-based test (Table 3); the A2, B1, and C1 German paper-based tests (Table 6); and the A2 and B2 German computer-based tests (Table 5) 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 (Table 3), the C1 level English computer-based test (Table 4), and the B2 level German paper-based test (Table 6).

Table 6

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.

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 (Table 7) 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.

Table 7

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

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 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.

5 Conclusion

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.

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 so, 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

Another limitation might emerge in connection with the data collection platform of the computer-based test. 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.

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.

6 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. 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.

In addition to 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.

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.

References

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

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.

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.

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

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

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

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.

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

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

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

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

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.

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

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.

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

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.

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.

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.
Ockey, G. (2007). Construct implication of including still image or video in computer-based listening tests. *Language Testing*, *24*, 517–537.
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.
Rost, M. (1990). *Listening in language learning.* New York, NY: Longman.
Sueyoshi, A., & Hardison, D. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*, 661–699.
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.
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.
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

Publications and conference presentations by the author

PUBLICATIONS

Dávid, G., Király, Zs., Kövér, Á., Mák, É., & Matuz, B. (2016). *Test-based listening exercises for the MA Language Development for Teachers' courses*. Budapest, HU: Eötvös Loránd Tudományegyetem.

Kövér, Á. (2018). Disciplinary differences of topical progression in discourse: The case of abstracts in applied linguistics and literary studies. *Working Papers in Language Pedagogy*, *12*, 15–26. Available at http://langped.elte.hu/WoPaLParticles/W12KoverA.pdf

Kövér, Á. (2018). Investigating an oral language proficiency examination: Analyzing the reliability of test scores with the many-facet Rasch measurement approach. *Konin Language Studies*, *6*(4), 505–520. Available at http://www.ksj.pwsz.konin.edu.pl/wp-content/uploads/2019/06/KSJ-64-505-520.pdf

Kövér, Á. (2017). Investigating listening comprehension skills: Empirical validation of test scores. *Working Papers in Language Pedagogy*, *11*, 80–95. Available at http://langped.elte.hu/WoPaLParticles/W11Kover.pdf

Kövér, Á., & Szűcs, Á. (2015). The motivation processes of MA in English Applied Linguistics students. *Working Papers in Language Pedagogy*, 9, 22–40. Available at http://langped.elte.hu/WoPaLParticles/W9Kover_Szucs.pdf

Szűcs, Á., & Kövér, Á. (2016). Reading skills involved in guided summary writing: A case study. *Working Papers in Language Pedagogy*, *10*, 56–72. Available at http://langped.elte.hu/WoPaLParticles/W10SzucsKover.pdf

CONFERENCE PRESENTATIONS

"Modifications Proposed in the Rating Scale for the M.A. in ELT Oral Language Examination at ELTE SEAS," National Scientific Students' Associations Conference (OTDK), Humanities Section, Language Pedagogy Subsection, Pázmány Péter Catholic University, Budapest, Hungary, April 8 – 10, 2015

"Revising Rating Scales of an M.A. in ELT Oral Language Examination," 25th IATEFL-Hungary Conference, Budapest, Hungary, October 9 –11, 2015

"The motivation processes of MA in English Applied Linguistics students," Poster Presentation, School of English and American Studies, Eötvös Loránd University, Budapest, Hungary, April 28, 2015