

Double play in listening assessment

Franz Holzknecht

This thesis is submitted in partial fulfilment of the requirements for the
degree of Doctor of Philosophy

Department of Linguistics and English Language

Lancaster University

October 2019

Abstract

All listening test developers are faced with a choice: whether to play the listening text once (single play), or twice (double play). Although practices of international high-stakes test providers vary widely in this regard, only a relatively small number of studies have investigated the effects of single and double play. The central focus of most studies has been whether double play had an effect on test takers' scores. Establishing effects on test scores is important, however an equally relevant question is whether double play impacts construct validity as conceptualised by Messick (1989, 1995). It has not been fully established in what ways double play influences test takers' response processes compared to single play and thus how it alters the listening construct.

The research in this thesis investigated the effects of double play on item parameters and test takers' response processes, their anxiety levels, and perceptions. In Study 1, 306 candidates responded to four listening tasks in a complex counter-balanced research design involving two conditions (single and double play), two task formats (multiple-choice and open format), and two questionnaires targeting listening strategies, test-taking strategies, anxiety levels, and test takers' perceptions. In Study 2, 16 candidates completed the same tasks in both conditions on an eye-tracker and performed verbal recalls, which were stimulated by their eye-movements while they had been solving the items. The verbal recalls were analysed in terms of candidates' cognitive processing, their use of listening strategies and test-taking strategies, and their anxiety levels.

The results from Study 1 confirm the common finding of previous research in that double play increases test scores. However, the results from Study 1 and Study 2 also agree in showing that double play is beneficial in terms of reducing construct-irrelevant variance and enhancing construct representation. Candidates displayed more higher-order cognitive processing and used a greater variety and a greater proportion of listening strategies in double play versus single play. Candidates also relied less on test-taking strategies and were markedly less anxious. In addition, items from both task formats were more reliable and showed better discrimination in the double play condition. The findings are discussed in light of the many competing priorities in listening assessment, including test purpose, authenticity, practicality, and construct validity.

Acknowledgements

I would like to thank all the people who have helped me complete this thesis: Luke Harding for offering advice, encouragement, and feedback throughout, Tineke Brunfaut and John Pill for reading and commenting on my confirmation panel documents, John Field for discussing my research with me, Rita Green for sharing her expertise in CTT and IRT, Michael Linacre for helping me with the Facets analysis, and Andrew Harris for the html programming. Many thanks to Carol Spöttl, who has been my mentor from the beginning and persuaded me to start this project. I would like to thank my colleagues Kathrin Eberharter, Benjamin Kremmel, and Matthias Zehentner for their help and support over the years. I am grateful to Doris Frötscher for giving me access to information on the tasks used in the research. I would also like to acknowledge all the students who participated in the project and the teachers who supported me.

Special thanks goes to my family: my daughters Luisa and Antonia, who kept me entertained and distracted whenever I did not work on the thesis, and my partner Eliza, without whom this would not have been possible.

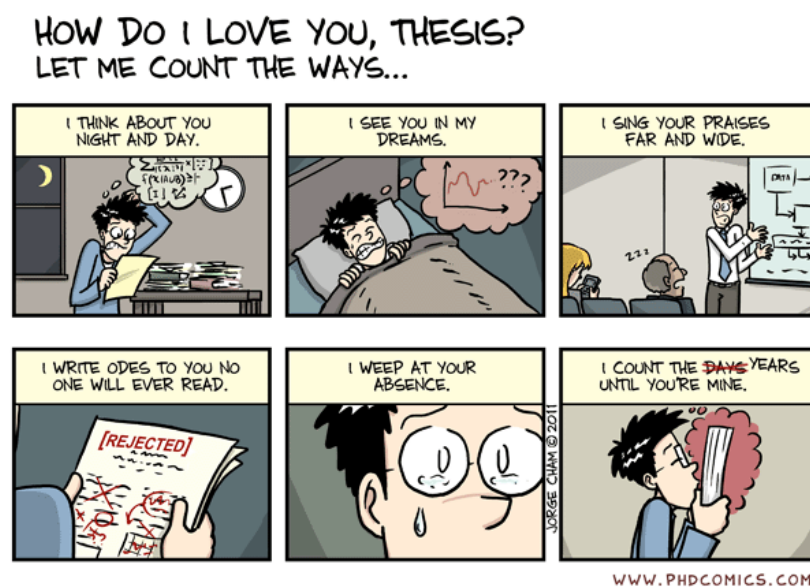


Table of contents

Abstract.....	ii
Acknowledgements	iii
Table of contents	iv
List of figures	viii
List of tables	x
1. Introduction	1
1.1. Background.....	1
1.2. Problem statement.....	4
1.3. Scope of the research	5
1.4. Terminology	7
1.5. Structure of the thesis	8
2. Literature review	10
2.1. Validity and validation	10
2.2. Response processes as part of validity evidence	13
2.3. Validity as framed in this thesis.....	14
2.4. Theoretical framework: responses processes in listening assessment	15
2.4.1. Cognitive processes.....	15
2.4.1.1. Rost 2011	15
2.4.1.2. Vandergrift and Goh 2012	17
2.4.1.3. Field 2013	18
2.4.1.4. Summary of the cognitive processing models of listening	19
2.4.1.5. Potential effects of double play on cognitive processing	20
2.4.2. Listening strategies.....	22
2.4.2.1. Potential effects of double play on the use of listening strategies	24
2.4.3. Test-taking strategies.....	26
2.4.3.1. Potential effects of double play on the use of test-taking strategies....	29
2.4.4. Anxiety	30
2.4.4.1. Potential effects of double play on anxiety	32
2.4.5. Summary of the theoretical framework.....	32
2.5. Previous research on double play in listening assessment	33
2.5.1. Studies looking at double play as part of a larger study.....	33
2.5.2. Studies focussing exclusively on double play	37

2.5.3. Summary of the findings on studies of double play	39
2.5.4. Field 2015	41
3. Methodology.....	45
3.1. Research questions	45
3.2. Methods to analyse test-related data	46
3.3. Methods to study test taker response processes in listening assessment	47
3.3.1. Questionnaires.....	48
3.3.2. Verbal protocols	50
3.4. Research design	53
3.4.1. Ethical issues.....	55
3.4.2. Research context	56
3.4.2.1. Matura listening tasks	57
3.4.2.2. Chosen tasks for the two studies	59
3.5. Study 1.....	61
3.5.1. Participants.....	62
3.5.2. Task adaptations	64
3.5.3. Questionnaire 1: strategies and anxiety	65
3.5.4. Questionnaire 2: biodata and task perception	66
3.5.5. Research design and procedure.....	67
3.5.6. Analysis	70
3.6. Study 2.....	71
3.6.1. Participants.....	72
3.6.2. Task adaptations	72
3.6.3. Research design and procedure.....	74
3.6.4. Prompts for verbal recall	76
3.6.5. Analysis	78
3.6.5.1. Transcription	78
3.6.5.2. Coding.....	80
3.6.5.3. Inter-coder reliability	85
4. Results Study 1.....	89
4.1. Classical test theory	89
4.2. Many-facet Rasch measurement	93
4.2.1. Single play versus double play	94
4.2.2. First play versus single play.....	96

4.2.3. First play versus second play	98
4.3. Answer change during the second play.....	101
4.3.1. Frequencies of answer change	102
4.3.2. Effects of answer change	103
4.4. Questionnaire 1: strategies and anxiety.....	106
4.4.1. Descriptive statistics	106
4.4.2. Factor analysis.....	109
4.4.3. Wilcoxon signed-rank test	112
4.5. Questionnaire 2: biodata and task perception.....	113
4.5.1. Topic familiarity.....	113
4.5.2. Task type familiarity	114
4.5.3. Perceived difficulty of the listening tasks.....	115
4.5.4. Face validity.....	118
4.5.5. Preference for single or double play	120
4.6. Summary of the main findings	122
5. Results Study 2.....	124
5.1. Cognitive processes.....	124
5.1.1. Examples from the data	124
5.1.2. Analysis of cognitive processes in single and double play	126
5.2. Listening strategies	133
5.2.1. Examples from the data	133
5.2.2. Analysis of listening strategies in single and double play.....	138
5.3. Test-taking strategies	143
5.3.1. Examples from the data.....	144
5.3.2. Analysis of test-taking strategies in single and double play.....	145
5.3.2.1. Test-management strategies	145
5.3.2.2. Test-wiseness strategies	149
5.4. Anxiety	151
5.4.1. Examples from the data	151
5.4.2. Analysis of anxiety in single and double play	152
5.5. Meta-commentary	154
5.5.1. Examples from the data	155
5.5.2. Analysis of meta-commentary	156
5.6. Summary of the main findings	157

6. Discussion	159
6.1. Convergence and summary of findings.....	159
6.1.1. Item and task statistics in single and double play	160
6.1.2. Cognitive processes in single and double play	161
6.1.3. Listening strategies in single and double play	163
6.1.4. Test-taking strategies in single and double play	164
6.1.5. Anxiety in single and double play.....	166
6.2. Connection with previous research.....	166
6.3. Extending the theory of listening assessment.....	171
6.4. Balancing priorities in listening assessment.....	173
7. Conclusion.....	178
7.1. Theoretical contribution	178
7.2. Methodological contribution	179
7.3. Practical contribution	180
7.4. Limitations.....	181
7.5. Suggestions for further research	182
7.6. Concluding remarks	184
Bibliography	185
Appendix.....	197
1. Ethical consent documents	197
2. Questionnaire 1	200
3. Questionnaire 2.....	201
4. Study 1: instructions for the test administration	202
5. Study 1: seating plan for the test administration.....	207
6. Study 1: test administration report.....	208
7. Study 2: double-coding document (excerpt)	209
8. Study 2: data excerpts	210
9. Study 1: exemplary Facets specifications file	215
10. Response frequencies for Questionnaire 2	216
11. Summary tables of observed response processes per participant	221

List of figures

Figure 1: Links in an interpretative argument (based on Kane et al., 1999, p. 9)	12
Figure 2: A processing model of listening based on Rost (2011), Vandergrift and Goh (2012), and Field (2013)	20
Figure 3: Cohen's definitions of test-taking strategies on a validity continuum (based on A. D. Cohen, 2011; and Doe & Fox, 2011)	29
Figure 4: Summary of the theoretical framework of response processes in listening assessment	33
Figure 5: Overview of the research design	55
Figure 6: Example of task instructions	65
Figure 7: Study 2: layout of MC tasks	73
Figure 8: Study 2: layout of NF tasks	73
Figure 9: Study 2: final coding categories	85
Figure 10: Study 2: coding categories for double-coding for coder 2 and coder 3	86
Figure 11: Study 1: Facets bias analysis and associated t-values between single play and double play across the two task types	95
Figure 12: Study 1: Facets bias analysis and associated t-values between single play and double play across the four tasks	96
Figure 13: Study 1: Facets bias analysis and associated t-values between the first play in double play and single play across the two task types	97
Figure 14: Study 1: Facets bias analysis and associated t-values between the first play in double play and single play across the four tasks	98
Figure 15: Study 1: Facets bias analysis and associated t-values between the first play and the second play in double play across the two task types	99
Figure 16: Study 1: Facets bias analysis and associated t-values between the first play and the second play in double play across the four tasks	100
Figure 17: Study 1: analysis process for answer changes during the second play	101
Figure 18: Study 1: coding categories and frequencies for the open question in Questionnaire 2 ("Why do you prefer double play?")	120
Figure 19: Study 2: cognitive processes while listening for single play and double play (all tasks)	128
Figure 20: Study 2: cognitive processes from the retrospective recalls and the post-listening stages for single play and double play (all tasks)	128
Figure 21: Study 2: cognitive processes while listening for single play and double play (MC and NF)	130
Figure 22: Study 2: cognitive processes while listening for single play and double play (MC1, MC2, NF1, NF2)	131
Figure 23: Study 2: total number of listening strategies as a proportion of overall metacognitive processing for single play and double play (all tasks; by stage of task completion)	139

Figure 24: Study 2: individual listening strategies as a proportion of overall metacognitive processing for single play and double play (all tasks; pre-listening only)	139
Figure 25: Study 2: individual listening strategies as a proportion of overall metacognitive processing for single play and double play (all tasks; while-listening only)	140
Figure 26: Study 2: individual listening strategies as a proportion of overall metacognitive processing for single play and double play (all tasks; post-listening only)	141
Figure 27: Study 2: total number of listening strategies as a proportion of overall metacognitive processing for single play and double play (MC and NF; by stage of task completion)	142
Figure 28: Study 2: total number of listening strategies as a proportion of overall metacognitive processing for single play and double play (MC1, MC2, NF1, NF2; while-listening only)	143
Figure 29: Study 2: test-management strategies as a proportion of overall metacognitive processing for single play and double play (all tasks; by stage of task completion)	146
Figure 30: Study 2: test-management strategies as a proportion of overall metacognitive processing for single play and double play (MC and NF; by stage of task completion)	148
Figure 31: Study 2: test-management strategies as a proportion of overall metacognitive processing for single play and double play (MC1, MC2, NF1, NF2; while-listening only)	149
Figure 32: Study 2: test-wiseness strategies as a proportion of overall metacognitive processing for single play and double play (all tasks; by stage of task completion)..	150
Figure 33: Study 2: test-wiseness strategies as a proportion of overall metacognitive processing for single play and double play (MC and NF; by stage of task completion)	151
Figure 34: Study 2: anxiety as a proportion of all coded quotations for single play and double play (all tasks; by stage of task completion, including retrospective recalls)	153
Figure 35: Study 2: anxiety as a proportion of all coded quotations for single play and double play (MC and NF; by stage of task completion, including retrospective recalls)	154

List of tables

Table 1: Test-management strategies as defined by Cohen versus listening strategies as defined by Vandergrift and Goh.....	27
Table 2: Condensed set of specifications of the Matura B2 listening exam for English	59
Table 3: Summary of the tasks used in the two studies	61
Table 4: Study 1: participants' age	63
Table 5: Study 1: participants' L1	63
Table 6: Study 1: participants with a second L1	63
Table 7: Study 1: research design.....	68
Table 8: Study 2: research design.....	75
Table 9: Study 2: separate data files and types of data for Participant 1	80
Table 10: Study 2: inter-coder agreement between the researcher and coder 2.....	88
Table 11: Study 2: inter-coder agreement between the researcher and coder 3.....	88
Table 12: Study 1: reliability, facility values, and discrimination indexes for the MC tasks in a double play condition and the NF tasks in a single play condition (sub-group 1).....	90
Table 13: Study 1: reliability, facility values, and discrimination indexes for the MC tasks in a single play condition and the NF tasks in a double play condition (sub-group 2).....	91
Table 14: Study 1: mean facility values of the MC tasks	92
Table 15: Study 1: mean facility values of the NF tasks	92
Table 16: Study 1: categorization of answer changes during the second play.....	102
Table 17: Study 1: frequencies of answer change during the second play across task types	103
Table 18: Study 1: frequencies and chi-square statistics for the answer change categories across the two task types	104
Table 19: Study 1: frequencies for the two NF specific answer change categories ...	106
Table 20: Study 1: descriptive statistics for the responses to Questionnaire 1: MC tasks in single and double play	107
Table 21: Study 1: descriptive statistics for the responses to Questionnaire 1: NF tasks in single and double play	108
Table 22: Study 1: KMO measure of sampling adequacy and Bartlett's test of sphericity for the factor analysis of the responses to Questionnaire 1	111
Table 23: Study 1: rotated factor matrix for the responses to Questionnaire 1.....	112
Table 24: Study 1: descriptive statistics for the three factors of the responses to Questionnaire 1.....	113
Table 25: Study 1: Wilcoxon signed-rank test and effect sizes for the three factors of the responses to Questionnaire 1	113

Table 26: Study 1: topic familiarity across two sub-groups and the four tasks	114
Table 27: Study 1: task type familiarity across the two sub-groups and the two task types	115
Table 28: Study 1: perceived difficulty of the listening tasks across the two sub-groups and the four tasks.....	116
Table 29: Study 1: perceived difficulty of the listening tasks for sub-group 1.....	116
Table 30: Study 1: crosstabulation of Wilcoxon signed-rank tests' p-values and effect sizes on perceived difficulty of the listening tasks for sub-group 1	117
Table 31: Study 1: perceived difficulty of the listening tasks for sub-group 2.....	117
Table 32: Study 1: crosstabulation of Wilcoxon signed-rank tests' p-values and effect sizes on perceived difficulty of the listening tasks for sub-group 2	117
Table 33: Study 1: face validity across the two sub-groups and the four tasks	118
Table 34: Study 1: face validity for sub-group 1.....	119
Table 35: Study 1: crosstabulation of Wilcoxon signed-rank tests' p-values and effect sizes on face validity of the listening tasks for sub-group 1	119
Table 36: Study 1: face validity for sub-group 2.....	119
Table 37: Study 1: crosstabulation of Wilcoxon signed-rank tests' p-values and effect sizes on face validity of the listening tasks for sub-group 2	119
Table 38: Types of data used to answer the research questions	160
Table 39: Threats to construct validity of a single play convention compared to a double play convention	173

1. Introduction

1.1. Background

A traditional convention in teaching and testing second language (L2) listening comprehension is repeating the listening text, so that students or test takers can listen to it twice (double play). Playing the listening text a second time is particularly common in teaching L2 listening in order to give learners the chance of adjusting to the characteristics of the recording, such as the speakers' accents, the voices in the recording, or the speech rate (Field, 2008, p. 159; Hubbard, 2017). As pointed out by Vandergrift and Goh, L2 listeners "often miss the first parts of an aural text and they struggle to construct the context and the meaning for the rest of the message" (Vandergrift & Goh, 2012, p. 4). Double play is therefore utilised in many national L2 school leaving exams, for example in the Standardised Austrian Matriculation Examination (Matura), the German Abitur (across all German federal states), or the English exam in the French Baccalauréat, where some of the English listening texts are even played three times. In addition, internationally recognised language tests such as the Cambridge English Assessment suite also feature double play in all of the listening tasks across the different levels.

Other international language test providers give test takers the choice over the amount of times the listening text is played. For example, in the current version of the British Council's Aptis Test candidates can choose whether they want to hear the listening text a second time by having control over the play button. Similarly, in the Oxford Test of English test takers can listen to the recording twice, but they can also move to the next task after only one hearing.

In a third group of international high-stakes listening tests, the texts are played twice for some of the tasks, but only once (single play) for other tasks. For example, two of the listening tasks in the current version of the widely-used Test of German as a Foreign Language (Test DaF) feature single play, while the third task utilises double play. Similarly, the Pearson Test of English (PTE) General uses a mix of single play and double play across the different levels, although the majority of listening texts are played only once. In the English examinations developed by Trinity College, on the

other hand, double play is the standard convention for most of the listening exams offered, and single play is only used for the highest level exam.

Finally, in a fourth group of international high-stakes tests, all of the listening texts are played only once. Widely-used English language assessments such as the Test of English for International Communication (TOEIC), the Test of English as a Foreign Language (TOEFL), the International English Language Testing System (IELTS), the PTE Academic, or the General English Proficiency Test (GEPT), all follow a single play convention throughout their listening assessments.

The tests outlined above are taken by millions of people every year for high-stakes purposes including immigration, university admission, or work-related decisions, however the reasons why language test providers choose double play over single play or vice versa are not always clear. This is particularly true for test providers utilizing double play, who generally do not provide a rationale as to why they play listening texts twice. More arguments can usually be found in support of single play. For example, in the case of the TOEFL the origins of the single play convention seem to date back to a study by Henning (1991), who found no positive effects of double play on item discrimination or on item response validity (as indicated by fit statistics) compared to single play, utilizing items from TOEFL tests conducted in the mid-1980s. Based on his findings Henning implied that there is no need to use double play in the TOEFL listening test, and today the test still features single play only. In other cases the preference for single play is most likely guided by practical considerations. If the listening texts are played only once, test providers can include a larger number of items in their tests, in the hope to thereby enhance reliability and construct representation (Fortune, 2004; Green, 2017, p. 44; Jones, 2011).

Another often mentioned argument for single play is that in most real-life situations people hear a listening text only once and thus listening assessments should also test this kind of “one-off” listening. According to Buck, listening in real life is characterised by processing information automatically (Buck, 2001, p. 172), so test takers should only be allowed one hearing in a listening test. If passages are played twice, so the argument goes, listening tests are not adhering to the important language testing principle of authenticity, that is “the degree of correspondence of the characteristics of a given language test task to the features of a TLU (target language use) task” (Bachman & Palmer, 1996, p. 23). While this seems to be true on the face of it, it can also be questioned whether people in real life always get only one chance at

understanding spoken language. As pointed out by Robin (2007), “[t]he notion that L2 learners must grab a flow of speech on the first try or lose the meaning is valid only for those events where the audio is not repeatable” (p. 110), and upon closer inspection it seems that situations where the audio is repeatable are more common than is generally assumed. For example, in face-to-face conversations people usually have the opportunity to ask for clarification should they miss or mishear information. In addition, considering the increasing role of technology in day-to-day interactions as well as in academic and professional domains, the notion that people only have one opportunity to understand meaning in real-life listening situations seems outdated (Hubbard, 2017, pp. 94–95; Robin, 2007, p. 110). Online conversations can be recorded, online videos can be paused and replayed, and academic courses are increasingly offered either fully online or in a hybrid form including online and offline content (Sun & Chen, 2016). In all of these scenarios, listeners have the chance to replay passages if they miss important information.

Even if it were true that single play is more authentic, it would not be a sufficient argument, as “just because a test task seems to be “authentic” does not mean that that test thus has construct validity” (Ockey & Wagner, 2018, p. 3). A task in a single play condition does not guarantee that the same type of listening will be performed as it would in real life because of the additional pressures associated with any type of language test, such as candidates’ use of test-taking strategies as well as their anxiety levels. It remains to be explored whether tasks in a single play condition are susceptible in the same way to these construct-irrelevant factors compared to tasks in a double play condition.

In sum, there is still much ambiguity on the question of single play versus double play in the current practices of test providers. Although repeatability seems increasingly common in real-life listening and double play is a widely accepted and applied method in teaching and learning L2 listening, many international test providers still adhere to a single play convention, or use a mixture between single play and double play in their listening tests. Thus, researching this issue also has the potential to create positive washback from testing practices to classroom practices, however, it is currently unclear how single play or double play affect listening comprehension.

1.2. Problem statement

As the above examples show, the question of single and double play remains controversial in the field of language testing (see also Buck, 2001, p. 172; and Field, 2008, p. 159), however, surprisingly few studies have investigated this issue. One reason for the lack of research in this area might be that studies on listening comprehension in general are still sparse. This can be vividly illustrated by quotations of leading listening researchers over the last 12 years, such as Vandergrift's observation in 2007 that despite the fact that "[l]istening comprehension lies at the heart of language learning, [...] it is the least understood and least researched skill" (2007, p. 191). More than a decade later, Buck echoed Vandergrift's concern, when he wrote that listening "is still the most neglected of the traditional four skills [, which] is unfortunate, because in a number of ways, [it] can be regarded as the most fundamental skill of all" (Buck, 2018, p. XI). Rost put it similarly by arguing that "we may just be scratching the surface of a deeper understanding of the fundamental processes and mechanisms that underpin our ability to communicate with members of our own species" (Rost, 2011, p. 1). These observations also hold true for the question of single play and double play in listening assessment, which has hardly been addressed by language testing researchers to date.

The dearth of published studies and continuing ambiguity in this area are particularly concerning in light of the differing practices of international high-stakes test providers outlined above. It could be argued that the decision to use single play instead of double play or vice versa could have a major impact on the validity of listening assessments, as it might not only effect takers' performance results, but also how they approach and complete a listening assessment task. Consequently, test takers' response processes could be influenced considerably by a single play or double play convention, and it is not yet fully known in what ways the mode chosen effects the listening construct. Test providers utilizing double play may therefore test a different construct than those who follow a single play convention.

Although a number of studies have investigated the effects of double play in L2 listening pedagogy and assessment, the great majority of these investigations looked at double play only as a secondary treatment as part of a larger investigation (see Section 2.5). In addition, the sole focus of most studies has been on whether double play impacts test scores. Most of the studies found that playing the recording a second time aided comprehension and increased test takers' scores (e.g. Lund, 1991), while a smaller

number of investigations reported that students did not benefit from double play as much as expected (e.g. Brindley & Slatyer, 2002).

However, to date only one study has investigated the effects of single play and double play on test takers' response processes (Field, 2015). This is arguably an equally relevant question for making decisions about language test design in comparison to the question of item difficulty, because test scores alone do not tell us anything about how test takers arrived at their answers. For this reason, test takers' response processes are seen as a crucial part of validity research (Hubley & Zumbo, 2017). To shed light on this important aspect, Field (2015) compared test takers' cognitive processing in the first play of a double play condition with their processing during the second play. He found that during the first play candidates heavily relied on word-level decoding and during the second play he observed increased levels of higher-order cognitive processes. Candidates also reported lower levels of anxiety during the second play compared to the first.

Although Field's research is important and insightful, it remains to be explored in what ways candidates' cognitive processes differ between a single play and a double play condition and whether double play impacts candidates' metacognitive processing (including their use of test-taking strategies). Therefore, extending and improving upon Field's (2015) study, the research presented in this thesis analyses different facets of double play which have not been investigated to date, in the hope that the results help us understand more fully how this convention affects item parameters as well as test takers' response processes and, consequently, the construct that is measured.

1.3. Scope of the research

The research is framed within modern validity theory and considers test takers' response processes as a crucial part of validity evidence. As pointed out by Messick, in order to establish construct validity, "possibly most illuminating of all [...] are direct probes and modeling of the processes underlying test responses" (Messick, 1995, p. 743). Instead of solely focussing on the product (i.e. the test score), it is vital to also investigate the test-taking process from the test takers' perspective to more fully understand the complexities of language assessment (Bachman, 1990, p. 269; Brindley & Slatyer, 2002, p. 390). Following Messick, Hubley and Zumbo argue that "[i]dentifying and understanding the mechanisms underlying how different respondents interact with, and

respond to, test items and tasks is essential to understanding score meaning and test score variation” (Hubley & Zumbo, 2017, p. 8).

The research in this thesis utilises innovative methods to broaden our understanding of how single play and double play affect construct validity. The research is designed to (1) test the replicability of previous studies’ findings on difficulty and to extend the scope of such quantitative findings by including a questionnaire on metacognitive processing and anxiety and another questionnaire on candidates’ perceptions (Study 1), and (2) to extend on these findings by seeking to understand the deeper construct-related dimensions of single versus double play (Study 2). In order to achieve this, a mixed methods research design was developed which yields multiple types of evidence. Mixed methods designs have been used increasingly in language testing in order to be able to triangulate findings and thereby gain clearer insights into the multi-faceted nature of assessment (Jang, Wagner, & Park, 2014).

In order to contrast findings with the main share of previous studies in this area, the first part of the research (Study 1) investigated in detail the effects of single play and double play on test scores and statistical item parameters such as item discrimination and reliability. By basing the study on a complex and counterbalanced design and controlling for potential confounding factors, it was hoped that the results would help us gain a fuller understanding of the impact of single play and double play on item characteristics compared to previous research. Several variables were carefully controlled for in the present study. First, the tasks used in the research were taken from past live papers of the Austrian Matura exam, a language test that is developed following best-practice guidelines and procedures. Potentially confounding factors such as task format, audio file length, number of items per task, difficulty of the items, standard setting results, or topics covered by the tasks, were taken into account in the research design. The tasks were also counterbalanced across the different conditions to avoid potential task ordering effects. In addition, two questionnaires were developed and administered alongside the tasks to tap into test takers’ metacognitive processing and anxiety levels as well as their perceptions of the tasks in the different conditions.

In the second part of the research (Study 2) eye-tracking was used in combination with retrospective and stimulated recalls to gain further insights into test takers’ cognitive processing, metacognitive processing, and anxiety levels in relation to single play and double play. Stimulating verbal recalls by replaying participants’ eye-movements is a promising method to gain insights into test takers’ thoughts and has the

advantage that the test-taking process does not need to be interrupted during the research (Brunfaut & McCray, 2015; Holzkecht et al., 2017; McCray, Alderson, & Brunfaut, 2012). As the same tasks were used in both parts of the research, the findings can be triangulated and compared across the different conditions.

Investigating these issues should not only inform listening test developers on the effects of single play and double play, but should also be insightful on a more general level. As pointed out by Taylor and Geranpayeh (2013a), “from a primarily cognitive perspective, the processes involved in second language listening are perhaps the least well described and analysed in the currently available literature on language assessment” (p.326). As the research study looked into cognitive processes, listening strategies, test-taking strategies, and anxiety levels of L2 listeners, the results will help to deepen our understanding of the listening construct.

1.4. Terminology

As some of the terminology in this thesis is used inconsistently and interchangeably in the research literature, the key terms are defined in this section. These terms appear throughout the thesis and are central to the research presented in the later sections.

Single play, as used in this thesis, refers to the convention of playing the listening text in a listening exercise or a listening assessment task only once. Test takers do not get a second chance to hear the text.

Double play refers to playing the listening text in a listening exercise or a listening assessment task twice. After the first listening, test takers can listen to the text a second time.

Response processes, as used in this thesis, are “the mechanisms that underlie what people do, think, or feel when interacting with, and responding to, the item or task and are responsible for generating observed test score variation” (Hubley & Zumbo, 2017, p. 2). In this thesis, response processes encompass cognitive processes, listening strategies, test-taking strategies, and anxiety. Each of these terms is defined in the following.

Cognitive processes are automatic mental actions which aid language comprehension (Rubin, 1981; Shiffrin & Schneider, 1977). They are one particular type of response process and operate in parallel, putting little or no strain on attentional load (Shiffrin & Schneider, 1977).

Listening strategies are another type of response process. In this thesis they are defined as goal-directed mental actions employed to aid listening comprehension (Afflerbach, Pearson, & Paris, 2008; Shiffrin & Schneider, 1977; Vandergrift & Goh, 2012). In contrast to cognitive processes, which operate automatically, listening strategies are consciously applied by the listener, thereby putting strain on attentional load. They usually operate in serial rather than in parallel (Shiffrin & Schneider, 1977).

Test-taking strategies are a third type of response process and consist of test-management strategies and test-wiseness strategies. **Test-management strategies** are controlled and goal-directed mental actions to find an answer to a test question. They are informed by both the test paper (the questions, answer options etc.) and the listening text. Similar to listening strategies, they put strain on attentional load. **Test-wiseness strategies**, on the other hand, while also being defined as controlled and goal-directed mental actions to find an answer to a question and putting strain on attentional load, are informed solely by the test paper (the questions, answer options etc.) or some other construct-irrelevant factor (e.g. guessing). Test-wiseness strategies are not informed by the listening text itself (A. D. Cohen, 2011; Doe & Fox, 2011).

1.5. Structure of the thesis

The thesis is comprised of seven chapters. Following the introductory chapter, Chapter 0 reviews relevant literature based on the identified research aims. The chapter first discusses validity and validation theory, focussing in particular on test takers' response processes as part of validity evidence. Next, relevant literature on four major groups of response processes in listening assessment are reviewed: cognitive processes, listening strategies, test-taking strategies, and anxiety. The four dimensions are discussed and it is considered how each of them could be impacted by single play versus double play. The chapter then turns to outline previous studies on double play in listening assessment and summarises the findings, with a detailed focus on the study by Field (2015), which is the only previous investigation on response processes in relation to double play.

Chapter 3 outlines the methodology used in the research. Based on the literature review, relevant research questions are formulated first. Next, the use of quantitative analyses, questionnaires, and verbal protocols stimulated by eye-movement recordings in validation studies are discussed, including a description of how these methods will be used in the research. The research design is then described in detail, comprising the

pilot studies conducted, the procedure for obtaining ethical consent, and the tasks used. As the research was framed in a mixed methods design consisting of two separate studies, the specific research design of the two individual studies is presented in turn, including the participants, additional materials used, as well as the analysis procedures.

The findings of the two individual studies are presented in Chapters 0 and 5. Chapter 0 outlines all of the findings in relation to Study 1, including results of two different statistical analyses of the test data (Classical Test Theory and Item Response Theory) and a detailed analysis of students' answer changes during the second play of the double play condition. The chapter also presents the results of the two questionnaires used in the research and outlines a summary of the main findings at the end. The results of Study 2 are then described in Chapter 5. Findings on the different groups of response processes are presented in turn, including illustrations of each of the coding categories from the verbal recall data as well as analyses of how the codes were applied across single play and double play. The chapter again concludes with a summary of the main findings.

Chapter 6 discusses the results of the research. First, the findings of the two individual studies are converged and summarised in relation to the five research questions. Then, the results are compared to past research on double play in L2 listening assessment, followed by a discussion of how the research in this thesis extends our current theory of listening assessment by considering the impact of single play versus double play on the construct that is measured. The chapter concludes with considering how the findings of the two studies relate to competing priorities in listening assessment, including test purpose, cognitive demand, reliability, and practicality.

Finally, in Chapter 7 the results of the research are summarised in terms of the theoretical, methodological, and practical contribution of the two studies. The chapter also outlines the limitations of the research and includes recommendations for future studies.

2. Literature review

Based on the identified research aims in the introduction, this chapter provides the scholarly backdrop of the thesis. In Section 2.1, validity and validation theory are discussed. A brief historical perspective on the notion of validity in language testing and testing more generally is presented first, followed by a discussion of two popular approaches to test validation and the different kinds of validity evidence they entail. One particular type of validity evidence – data on test takers’ response processes – is discussed in more detail in Section 2.2, as this is the main focus of the research in the thesis. The section outlines a definition of response processes and identifies their role in validation research.

Next, the chapter reviews relevant response processes for listening assessment in Section 2.4. The section discusses four main groups of response processes in detail – cognitive processes, listening strategies, test-taking strategies, and anxiety – and presents them as the theoretical validation framework of the thesis. As the thesis aims to identify how double play affects these four dimensions, the section also considers how double play could potentially impact each of them.

Section 2.5 then turns to consider previous studies on double play in listening assessment. The section discusses studies which have looked at double play as a secondary part in a larger study, as well as research which has focused exclusively on double play. Given the relatively small literature on this topic, the results of the individual studies are outlined in detail first before a synthesis of the findings is presented. Particular attention is then given to the study by Field (2015), as it is the only one focussing on response processes in relation to double play to date.

2.1. Validity and validation

Validity is the central concern in all assessment. In simple terms, “[a] test is said to be valid to the extent that it measures what it is supposed to measure” (Henning, 1987, p. 89). This basic definition, which frames validity as an “all-or-nothing attribute” (Chapelle, 1999, p. 255), dates back to the classic work by Lado (1961). However, the way in which validity is conceptualised has changed over the years.

Early conceptions of validity consisted of a number of distinct and independent types, the most prominent of which were content validity, criterion-related validity, and

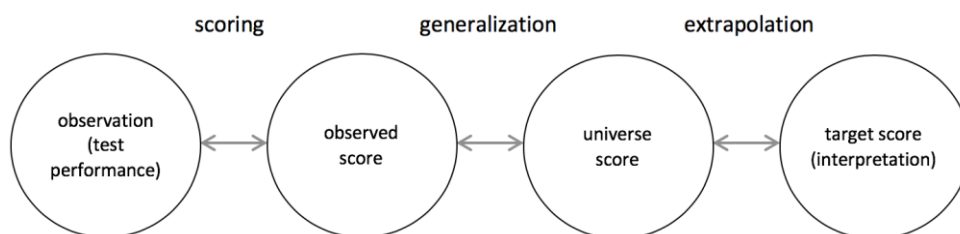
construct validity (Cronbach & Meehl, 1955). Content validity is related to the representativeness of the test content and how it corresponds to the test specifications (Hughes, 2003). It has traditionally been established through a systematic collection of expert judgements. Criterion-related validity, sometimes also referred to as external validity (Alderson, Clapham, & Wall, 1995), concerns the strength of agreement of a candidate's test score between two independent tests measuring the same trait. If high correlations between test scores were found, a test was seen as being valid (see Oller, 1979, pp. 417–418). Finally, construct validity was established through empirically investigating the extent to which test scores are an accurate representation of the theoretical expectations of the trait that was measured (Cronbach & Meehl, 1955). Over the years a number of additional types of validity were added, such as face validity and response validity, but the theoretical division into different 'validities' was maintained (Chapelle, 1999).

This changed in the late 1980s and early 1990s, most notably through the seminal work of Messick (1989), who framed validity as a unitary concept with construct validity at its core. Instead of referring to different types of validity, Messick argued that test interpretation and use should be the main object of validation efforts, and construct validity should be seen as central. He defined validity as “an overall evaluative judgment of the degree to which evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores” (Messick, 1989, p. 13). This position was supported by Bachman (1990), who also emphasised that validity should be concerned with test consequences, however Bachman explicitly separated consequences from the wider implications of test use. Thus, since the early 1990s, the standard view within the field of language assessment, and assessment more generally, has been that “validation research should consist of gathering evidence about the meaning of test scores” (Chapelle & Voss, 2013, p. 4). However, exactly how one goes about gathering evidence to establish validity is not straightforward, and Messick's influential work has been criticised for failing to offer practical guidance to test developers (see, for example, Bachman, 2005; Chapelle, Enright, & Jamieson, 2008; Davies & Elder, 2005; Kane, Crooks, & Cohen, 1999; Xi, 2008).

The need for concrete validation practices was first comprehensively addressed by Kane, who developed the argument-based approach to test validation (Kane, 1992, 2001, 2012). The general idea of Kane's model, which itself is based on the work by

Toulmin (1958), is that data needs to be gathered to find evidence for claims about test score interpretation. This set of claims, warrants, and assumptions forms a chain of inferences which provide meaning to test performances (Figure 1). The chain comprises three links, from scoring (e.g. the reliability and validity of scoring criteria and procedures, authenticity of testing conditions etc.), to generalization (e.g. the reliability and representativeness of the tasks used), to extrapolation (e.g. the targeted cognitive processes and how similar these are to real life processes). Each of the links needs to be backed up by sufficient evidence to establish a convincing validity argument. Bachman (2005) draws on the argument-based approach and considers it “an important move in language testing away from the highly abstract unified model of validity” (Bachman, 2005, p. 17). Kane’s framework has been applied in a number of studies in the field of language testing (see, for example, Chapelle et al., 2008; Cheng & Sun, 2015; Enright & Quinlan, 2010; Frost, Elder, & Wigglesworth, 2012; Knoch & Chapelle, 2018; Youn, 2015) and has been described as being more structured, more logical, more specific, and more productive than other approaches to validation (Chapelle, Enright, & Jamieson, 2010).

Figure 1: Links in an interpretative argument (based on Kane et al., 1999, p. 9)



A different model of test validation, which has been particularly popular in the British and European context, is Weir’s socio-cognitive framework (O’Sullivan & Weir, 2011; Weir, 2005). Weir proposed two different types of validity evidence, *a priori* and *a posteriori*, and divided each into a number of sub-categories. *A priori* evidence needs to be gathered before the test is administered by clearly defining the cognitive construct that is measured – referred to as “theory-based validity” or “cognitive validity” (Glaser, 1991). In addition, *a priori* validation consists of investigating the extent to which the test tasks correspond to the real world – referred to as “context validity”. *A posteriori* evidence, on the other hand, is collected after the test has been administered and encompasses “scoring validity” (the extent to which test

performances are scored reliably and validly), “criterion-related validity” (the correlation of test scores with those of other assessments, see discussion above), and “consequential validity”, which scrutinises the consequences of test outcomes on test takers and society more generally. Weir’s framework has been prominently applied by Cambridge English Language Assessment (see, for example, Taylor & Geranpayeh, 2013b for the application of the framework in the context of listening). However, it has also been criticised by proponents of the argument-based model, in that “the approach of defining multiple types of validities runs contrary to Messick’s (1989) statement about validity as a unitary concept” (Knoch & Chapelle, 2018, pp. 479–480).

2.2. Response processes as part of validity evidence

One of the five main sources of validity evidence in the authoritative *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) is the collection of data on test takers’ response processes. Response processes, as defined by Hubley and Zumbo (2017), are “the mechanisms that underlie what people do, think, or feel when interacting with, and responding to, the item or task and are responsible for generating observed test score variation” (Hubley & Zumbo, 2017, p. 2). Response processes can be placed within both validation models discussed in the last section. In Kane’s argument-based framework, evidence of representative response processes falls under the link of extrapolation. In terms of Weir’s approach to test validation, response processes are reminiscent of what is subsumed under the notion of “cognitive validity” (see also Field, 2013). However, in addition to the focus on test takers’ cognitive processes for establishing validity as suggested by Weir (2005) and Field (2013), Hubley and Zumbo also explicitly include test takers’ “emotions, motivations, and behaviours” (Hubley & Zumbo, 2017, p. 2) in their definition of response processes, although it is not entirely clear what the authors mean by “behaviours”.

Collecting evidence of response processes as part of establishing a test’s validity has traditionally been neglected across the social, behavioural, and health sciences (Padilla & Benítez, 2014) and, despite a recent surge of studies in this area, is still underrepresented in validation research (Hubley & Zumbo, 2017, pp. 7–8). This is surprising, since Messick himself wrote that in order to establish construct validity, “possibly most illuminating of all [...] are direct probes and modeling of the processes underlying test responses” (Messick, 1995, p. 743). Within the field of language

assessment, the traditional focus in validation research has also been on quantitative methods (see Lumley & Brown, 2005). Although introspective studies have been conducted since the early 1990s (e.g. Buck, 1990), they have only recently come to be regarded as integral components of validity research. Recent years have seen an influx of research on test takers' response processes (Sasaki, 2013) and nowadays introspective methods are seen as an indispensable tool in validation research, particularly in mixed methods studies (Turner, 2013).

There are a variety of research techniques to investigate response processes. Referring to the *Standards* (AERA et al., 2014), Messick (1989), and Padilla and Benítez (2014), Hubley and Zumbo (2017) list a number of relatively novel methods alongside the widely used verbal protocols and cognitive interviews. These new methods have been employed increasingly over the last years across disciplines and include, among others, eye-tracking, tracking response development (e.g. by analysing the changing of responses), or statistical models to infer test takers' response processes.

2.3. Validity as framed in this thesis

In this thesis, validity will be framed in Messick's (1989, 1995) terms. That is, the thesis seeks to "[gather] evidence about the meaning of test scores" (Chapelle & Voss, 2013, p. 4) in order to infer how single play and double play impact the validity of listening assessment instruments. The main aim of the research is to investigate in what ways single and double play influence test takers' response processes, to draw conclusions about two major threats for construct validity: construct-irrelevant variance and construct underrepresentation (Messick, 1989). Construct-irrelevant variance relates to factors outside the construct leading to success on a test, such as the use of background knowledge instead of language skills, the use of test-wiseness, or feelings of anxiety. Construct underrepresentation concerns a test's failure to measure parts of the construct and can occur when assessment instruments are too narrow. The thesis will focus on these two aspects of construct validity and in so doing will delineate recommendations for listening test development.

2.4. Theoretical framework: responses processes in listening assessment

In general, the construct of interest in any listening test are cognitive processes and metacognitive strategies leading to successful comprehension. In order to allow for meaningful decisions based on test scores, listening tests should target cognitive processes and metacognitive strategies which are close to those that test takers need to master in real-life listening. However, as any test is an artificial situation, other factors irrelevant to the construct also play a role. According to Golchi (2012), construct-irrelevant test-taking strategies and anxiety may account for the total of construct-irrelevant variance in listening tests, that is assuming that the listening tests are well-designed, valid tests, and there are no other construct-irrelevant factors such as scoring issues, other skills being tested through response format, working memory effects, attention regulation factors, test motivation, or cheating. As it would be beyond the scope of this thesis to investigate all of these areas, the research will focus on test-taking strategies and anxiety as construct-irrelevant factors. The aims of this section are twofold: First, the four response processes (cognitive processes, metacognitive strategies, test-taking strategies, and anxiety) will be discussed in turn; and second, by doing so, it will be considered how double play could potentially influence each of them.

2.4.1. Cognitive processes

Cognitive processes, as used in this thesis, are automatic mental actions which aid comprehension (Rubin, 1981; Shiffrin & Schneider, 1977). Researchers have developed different cognitive processing models of listening over the years. The models most often cited in the literature on L2 teaching and testing will be outlined in the following and it will be considered how double play in listening assessment might influence the processes discussed in the different models.

2.4.1.1. Rost 2011

Rost's (2011) model of listening comprehension consists of four different layers of processing: neurological processing, linguistic processing, semantic processing, and pragmatic processing. These layers of processing overlap and complement each other.

The first, neurological processing, describes the neurological activities that underlie all other types of processing. It includes the mechanisms of the ear, the transduction of the mechanical sound signals to neural activity, the stimulation of the auditory cortex by neural activity, and the subsequent translation of the neural signal by the different areas of the brain. Although the basic neurological processes underlying our ability to listen are the same for everyone, individuals control these functions very differently. Rost argues that it is likely that individuals display differences in all neural mechanisms.

The next stage in Rost's model of listening is linguistic processing, which he defines in bottom up fashion. According to Rost, when the speech signal reaches the brain, listeners first identify units of spoken language (referred to as intonation units or pause units), which are stored in short-term memory and then processed further. The next step is the recognition of words and lexical phrases, which happens automatically for advanced listeners. This step alone is very complex, as listeners do not simply recognise the meaning of a word, but at same time activate knowledge such as the word's lemma and its collocations, or available phonotactic knowledge about the language (for example knowledge about consonant assimilation, cluster reduction or dropping, or vowel changes). Word recognition is then followed by syntactic parsing, i.e. situating words within an utterance by applying grammatical knowledge. Rost differentiates between two overlapping and converging levels of syntactic parsing: making sense of only a sentence or an utterance, and deciphering the meaning of the discourse as a whole. Syntactic parsing is aided by pragmatic and intertextual knowledge, and familiarity with formulaic sequences and semantic roles. The last part of the linguistic processing stage described in Rost's model is the integration of non-verbal cues, such as gestures or eye and face movements, which help listeners confirm meaning.

The third layer in Rost's model of listening is semantic processing, which subsumes the concepts of comprehension, schemata, inferencing, and memory. Thus, in contrast to linguistic processing, which is conceptualised in bottom up fashion (starting at the speech signal), semantic processing is described by Rost as top down (starting with the listener's memory). Comprehension, according to Rost, "is the experience of understanding what the language heard refers to in one's experience or in the outside world" (p. 54). In order to comprehend the input, listeners constantly need to check if the incoming information is new (not active in memory) or given (already

active), and they are likely to give priority to information which is new. Listeners are aided in their understanding by intonational cues, i.e. when speakers put stress on information that is new, and by the activation of schemata. In order to comprehend successfully, listeners also regularly need to infer the meaning of utterances through reasoning. In addition, due to limited short-term memory capacities, listeners often rely on compensatory strategies such as skipping or approximating incoming information.

The final layer in Rost's model is pragmatic processing. Pragmatic processing is concerned with the social and contextual dimensions of the listening situation. In listening tests pragmatic processing is necessarily limited, as test takers are not active participants in the act of communication. Essential parts in Rost's model at this stage are the norms of one-to-one social interaction, affective involvement in conversations, shaping the interaction with responses, and connecting with the interlocutor verbally and non-verbally. However, these do not usually play a role in listening test situations. Instead, in many testing contexts these processes are assessed in speaking exams. Still, listening test takers need to display some degree of pragmatic processing, such as inferring the speaker's intentions from the context of the speech situation. They also need to be able to detect when conversational maxims are violated (for example when a speaker is being ironic), and to interpret a speaker's tone and emotions.

2.4.1.2. Vandergrift and Goh 2012

Vandergrift and Goh's (2012) model of listening comprehension is similar to Rost's outlined above, in that it integrates the different types of processes (linguistic, semantic, and pragmatic) into one coherent framework. However, Vandergrift and Goh base their model on the work by Levelt (1989, 1993, 1995), whose main focus was to explain speech production, and Anderson (1995), who proposed different levels of processing for listening comprehension. They therefore incorporate both speaking and listening processes into one holistic framework in order to account for both one-way and interactive listening. As the focus of this thesis is on testing listening in one-way settings, however, the speech production side of the model will not be discussed in detail here. Vandergrift and Goh point out that theirs is "only a working model" (2012, p. 38), as it does not account for the affective dimension of the listening process.

The listening comprehension side of their model is divided according to the three processing stages posited by Anderson (1995): perception¹, parsing, and utilisation. These can be seen as synonymous to Rost's "linguistic processing". The first one, perception, involves "the recognition of sound signals by the listener as words or meaningful chunks of language" (Anderson, 1995; as cited in Vandergrift & Goh, 2012, p. 41). This is achieved by the separation of language-relevant sounds from irrelevant sounds by the "acoustic-phonetic processor", and becomes automatic with increasing proficiency.

In the next stage the "parser" takes over from the "acoustic-phonetic processor". The "parser" assigns meanings to individual words by activating lexical knowledge such as lemmas and lexemes, and situates these words within a sentence by applying syntactic knowledge. Similar to the perception stage, parsing operates in a bottom up manner (from acoustic signals to word meanings), but is also informed by top down processing from the final stage in the listening model.

This final stage is termed "the conceptualiser". It is at this stage when individual clauses or sentences are connected to interpret meaning. This can happen at micro-level (interpreting the meaning of single utterances or parts of utterances) and at macro-level (interpreting the meaning of entire conversations). The conceptualiser is informed by the listeners' pragmatic knowledge (which Rost includes as a separate layer in his model, as described above), their discourse and background knowledge (the application of which are subsumed by Rost under "semantic processing"), and by the listeners' goals.

Vandergrift and Goh echo Rost's observation that all of these processes happen in parallel and constantly inform each other. For example, as listeners start to understand the meaning of a text (in the utilisation stage), they will find it easier to identify the meanings of individual words (in the parsing stage), and vice versa. As listeners' proficiency increases, their parallel processing will become more effective and accurate.

2.4.1.3. Field 2013

One of the most recent models of listening comprehension processes was proposed by Field (2013). Like Vandergrift and Goh, Field bases his model on Anderson's (1995) three processing stages. However, based on the work by Cutler and Clifton (1999), he

¹ In the 2000 edition of his book Anderson changed the name of this processing stage to "decoding".

subdivides Anderson's first stage (perception) into two separate operations. Field also splits up the final stage of Anderson's framework (utilisation), resulting in a bottom-up model consisting of five types of processing: input decoding, lexical search, syntactic parsing, meaning construction, and discourse construction. Like Rost (2011) and Vandergrift and Goh (2012), Field stresses that listeners operate all of these processes in parallel, rather than strictly hierarchical as the model might suggest.

The first three stages in Field's model are defined as lower-level processes and include input decoding, lexical search, and parsing. Lower-level processes "take place when a message is being encoded into language" (Field, 2013, p. 96). As such, when an acoustic signal reaches the ear, listeners first decode the input in terms of whether or not it is speech, and if it is identified as speech, they divide it into phonological segments by applying phonological knowledge. The next stage of lower level processing is lexical search, which involves the recognition of individual words by applying lexical knowledge. These words are then put into a syntactic pattern at clause level by applying syntactic knowledge in the parsing stage. The output of lower-level processing is the bare meaning of an utterance at clause or sentence level (the "proposition").

In contrast to lower-level processes, through which listeners identify the literal meaning of messages, higher-level processes "are those associated with building meaning" (Field, 2013, p. 96). Field divides higher level-processes into meaning construction and discourse construction. Meaning construction is concerned with relating the literal meaning of utterances or sentences to the context in which they occurred, by applying pragmatic, external (or world) and discourse knowledge (this is parallel to the knowledge sources in Vandergrift and Goh's model). Finally, discourse construction takes the meaning of the message further by relating it to the speech event as a whole, again by applying external knowledge. These two higher-level processes can thus be seen synonymous to what Vandergrift and Goh (2012) define as micro-level and macro-level conceptualizations in the last stage of their model.

2.4.1.4. Summary of the cognitive processing models of listening

Before considering how double play could potentially impact listening comprehension processes in terms of the three different models outlined above, it is necessary to discuss the similarities and differences between the models. It clearly emerged from the description of the three models that there are a number of parallels between them. This

is especially true for the models by Vandergrift and Goh (2012) and Field (2013), as both of them are based on Anderson's (1995) three processing stages. In contrast, Rost's (2011) four processing layers follow a conceptually different division, and neurological processing as defined by Rost is not included in Vandergrift and Goh's and Field's models. However, Rost's linguistic and semantic processing layers can be seen as umbrella terms subsuming the various processing stages and knowledge sources described by Vandergrift and Goh and Field. Especially linguistic processing, which is described by Rost in bottom up manner, parallels the different stages of the Anderson approach. Similarly, semantic processing as defined by Rost is heavily influenced by external knowledge, and can therefore be seen as relating to the different knowledge sources identified by Vandergrift and Goh and Field. These knowledge sources in turn influence the different processing stages in top down fashion. For the final layer in Rost's model, pragmatic processing, only the areas relevant for one-way listening will be considered in this discussion. As such, Rost's pragmatic layer comes into play at higher level listening processes. Based on this reasoning, Figure 2 is an attempt to summarise the three different models, accounting for the conceptual differences as well as the parallels between them.

Figure 2: A processing model of listening based on Rost (2011), Vandergrift and Goh (2012), and Field (2013)

Processing layers				Knowledge sources	
Rost, 2011		Vandergrift and Goh, 2012		Field, 2013	
semantic ↓	conceptualiser	macro	discourse construction	higher	external (text type)
		micro	meaning construction		external (world, speaker)
					external (world, speaker, topic)
linguistic ↑	parser	acoustic-phonetic processor	parsing	lower	pragmatic
			lexical search		syntactic
		input decoding	phonological		
neurological →					

2.4.1.5. Potential effects of double play on cognitive processing

In this section it will be discussed how double play might impact the different layers and stages of Figure 2. The main structure for this discussion will be Rost's four layers

of processing, however, Vandergrift and Goh's and Field's terminology will be used in the appropriate sections. The discussion will start at the bottom of the figure with neurological processing, and will then consider linguistic, semantic, and pragmatic processing in turn.

In terms of neurological processing, it can be argued that hearing a text twice as opposed to only once might impact certain underlying neurological functions. Especially the notions of consciousness and attention, which Rost discusses at the neurological stage of his processing model, are relevant to testing listening and could potentially be affected by double play. Consciousness, according to Rost, "guides the person's intentions to experience the speaker's world and to attempt to construct meaning from this experience". Attention, on the other hand, is "the operational aspect of consciousness" (Rost, 2011, p. 21). As attention has limited capacity and can be directed selectively while listening, it seems reasonable to hypothesise that test takers would focus on different parts of the listening text or listening task depending on whether they hear the recording once or twice.

Double play could potentially also impact test takers' linguistic processes. In terms of lower level linguistic processes such as input decoding or lexical search, test takers might decode the meaning of a particular word only during the second play. This in turn could influence the parsing stage, i.e. how test takers understand the utterance. Moving up the processing stages of Vandergrift and Goh and Field, test takers' understanding of the utterance would further impact both micro and macro conceptualising, or to use Field's terminology, both meaning construction and discourse construction. Similarly, if test takers know that they will hear the text twice, they might only listen to details during the first play, thus operating at lower level linguistic processing, but more globally during the second play, thereby employing higher level linguistic processes, or vice versa.

Double play might also influence semantic processing, which according to Rost relates to how listeners use their knowledge to aid comprehension. For example, Rost mentions the strategy of "incompletion" at this stage of his model; i.e. "maintaining an incomplete proposition in memory [and] waiting until clarification can be obtained" (Rost, 2011, p. 70). This strategy would be particularly important for L2 learners. It could be argued that listening exams where the text is played twice are more likely to assess this strategy, as clarification might often only be obtained during the second play. In addition, the different knowledge sources included in Vandergrift and Goh's and

Field's models, which relate to semantic processing as defined by Rost, might also be impacted by listening twice. For example, knowledge of discourse structure or text type, which test takers can only benefit from after the first play, might aid them in their higher-level processing during the second play.

Another area relevant to semantic processing, which was investigated in relation to listening comprehension to some extent in the 1980s and 1990s, but has received relatively little attention from listening researchers in recent years, is schema theory. In his seminal work on schema theory, Bartlett (1932) determined that language comprehension is heavily influenced by knowledge of different cognitive structures, such as topics, text types, or situations (so-called schemata). Since then, it has been established by a number of studies that schemata aid L2 listening comprehension (see for example Chang & Read, 2006; Chiang & Dunkel, 1992; Long, 1990; Markham & Latham, 1987; Schmidt-Rinehart, 1994; Teng, 1998). Although the majority of these investigations used double play in their methodology, none of them addressed in detail whether repeating the input had any effect on the results in terms of enhancing schemata. From a validity point of view, it is relevant to know whether listening test takers exposed to unfamiliar topics develop content schemata during their first listening of a passage, and access these during the second listening.

Finally, double play could also influence test takers' pragmatic processing in listening tests. Interpreting a speaker's tone and emotions, or inferring the speaker's intentions from the context, might be challenging for L2 learners in single play listening tasks. Listeners might be taxed enough at a cognitive level developing a propositional meaning on the first listening, and the second listening might allow them the opportunity to test any hypotheses about pragmatic meaning.

It has emerged from this discussion that replaying the text in listening tests could have an effect on the cognitive processes test takers employ in order to comprehend the input. It was argued that double play could potentially affect all processing stages of the three listening models outlined above.

2.4.2. Listening strategies

In contrast to cognitive processes, which operate automatically, listening strategies are conscious, goal-directed mental actions to aid comprehension (A. D. Cohen & Upton, 2007; Shiffrin & Schneider, 1977). Listening strategies are particularly important for

L2 learners, as their cognitive processes might not have been developed to the extent that comprehension is fully fluent (Faerch & Kasper, 1986). Therefore, any model of L2 listening comprehension needs to take listening strategies into account.

One of the most comprehensive models of listening strategies has been developed by Vandergrift and Goh (2012), who conceptualise listening strategies within a framework of metacognition. Metacognition, which is often described as thinking about thinking (Flavell, 1976), plays a crucial role in successful language learning and L2 comprehension (Wenden, 1987). In Vandergrift and Goh's framework for listening, metacognition is separated into 1) metacognitive knowledge, which is the knowledge about one's own personality, the task at hand, and strategies which might be effective to complete the task; 2) metacognitive experience, which relates to prior experience with and use of effective strategies; and 3) strategy use, which is the conscious application of strategic knowledge to aid understanding. Based on the work by Goh (1998, 2002), O'Malley and Chamot (1990), Oxford (1990), Vandergrift (1997, 2003), and Young (1997), Vandergrift and Goh propose the following listening strategies, each of which are further split into separate sub-categories not described here:

1. Planning: Developing awareness of what needs to be done to accomplish a listening task, developing an appropriate action plan and/or appropriate contingency plans to overcome difficulties that may interfere with successful completion of a task.
2. Focusing attention: Avoiding distractions and heeding the auditory input in different ways, or keeping to a plan for listening development.
3. Monitoring: Checking, verifying, or correcting one's comprehension or performance in the course of a task.
4. Evaluation: Checking the outcomes of listening comprehension or a listening plan against an internal or an external measure of completeness, reasonableness, and accuracy.
5. Inferencing: Using information within the text or conversational context to guess the meanings of unfamiliar language items associated with a listening task, to predict content and outcomes, or to fill in missing information.
6. Elaboration: Using prior knowledge from outside the text or conversational context and relating it to knowledge gained from the text or conversation in order to embellish one's interpretation of the text.

7. Prediction: Anticipating the contents and the message of what one is going to hear.
8. Contextualization: Placing what is heard in a specific context in order to prepare for listening or assist comprehension.
9. Reorganizing: Transferring what one has processed into forms that help understanding, storage, and retrieval.
10. Using linguistic and learning resources: Relying on one's knowledge of the first language or additional languages to make sense of what is heard, or consulting learning resources after listening.
11. Cooperation: Working with others to get help on improving comprehension, language use, and learning.
12. Managing emotions: Keeping track of one's feelings and not allowing negative ones to influence attitudes and behaviors.

(Vandergrift & Goh, 2012, pp. 277–284)

Some of these strategies do not usually play a role in listening assessment. For example, using learning resources (part of strategy 10 in the list above) may be a useful strategy for classroom practice, but would not be relevant in listening tests, as candidates generally are not allowed to use aids such as dictionaries. In addition, the strategy of cooperation (strategy 11 in the list above) would not be applicable to assessing listening, as listening tests are usually completed alone. Thus, overall, 11 of the 12 listening strategies proposed by Vandergrift and Goh (with only the linguistic part of strategy 10 in the list above) should ideally be assessed in L2 listening tests. In the following, it will be discussed how the use of these strategies may be influenced by double play in listening assessment.

2.4.2.1. Potential effects of double play on the use of listening strategies

Double play could potentially impact the use of all of the strategies relevant to listening assessment proposed by Vandergrift and Goh discussed above. Starting from the top of the list with the strategy of planning (strategy 1), students would quite likely plan their completion of a listening task differently if they know from the outset that they are going to hear the recording twice. They might plan to listen on a more general level

during the first play and pay more attention to specific questions during the second play, or vice versa.

Similarly, students may focus their attention differently as well (strategy 2). In a double play situation, they might pay selective attention to unanswered questions during the second play, whereas in a single play situation, they might try to maintain their attention throughout as they know they will not get a second chance.

The strategies of monitoring and evaluation (strategies 3 and 4) may play a bigger role in tasks which utilise double play. That is because students simply have more opportunities to monitor and evaluate their comprehension and performance on a task if they hear a recording a second time. In a single play condition L2 learners' limited processing capacity might already be at the limit through the use of other strategies such as focussing attention, and monitoring and evaluation might therefore play a smaller role.

On the other hand, the next three strategies in Vandergrift and Goh's model - inferencing, elaboration, and prediction (strategies 5, 6, and 7) - could be more important in a single play condition. This is because students may miss important information during the first play and in the absence of a second play try to infer what they have missed from different parts of the listening text, or elaborate on the missed information from outside sources, such as personal experience or world knowledge. Similarly, if students know that they are going to hear a listening text only once, they might put more cognitive resources into predicting what they are going to hear.

Contextualisation and reorganising (strategies 8 and 9), in contrast, could potentially be used more often in double play situations. When students get the chance to hear a listening text again, they might be able to better contextualise what they hear during the second play as they are already familiar with the recording. They might also have more opportunities to reorganise what they have heard and thereby increase comprehension, for example by making a mental summary after the first play and using it during the second play.

Finally, using linguistic resources and managing emotions (strategies 10 and 12) could again be more relevant for test takers if they hear the listening text only once. Test takers might rely more on simple linguistic strategies such as translation in single play, as their limited processing capacity may not allow them to utilise more resource-intensive strategies such as monitoring or evaluation. In addition, they may be more anxious if

they know that they are going to hear a text only once, therefore relying more on strategies which help them keep their emotions in check.

In summary, the use of all of the listening strategies relevant to listening assessment proposed by Vandergrift and Goh could potentially be impacted by double play. However, no study has yet investigated in detail how double play impacts the use of listening strategies.

2.4.3. Test-taking strategies

A cognitive model of L2 listening assessment not only needs to include cognitive listening processes and listening strategies, but also processes which are specific to the test situation. One of the most frequently cited authors in this strand of research within the field of language testing is A. D. Cohen, who uses the umbrella term “test-taking strategies” to refer to “the consciously selected processes that the respondents [use] for dealing with both the language issues and the item-response demands in the test-taking tasks at hand” (A. D. Cohen, 2006, p. 308). Cohen separates test-taking strategies into three distinct types: language learner strategies, test-management strategies, and test-wiseness strategies.

The first one, language learner strategies, is defined by Cohen as the learners’ operationalization of “their basic skills of listening, speaking, reading, and writing as well as the related skills of vocabulary learning, grammar, and translation” (A. D. Cohen, 2006, p. 308). Therefore, in the context of listening, what Cohen subsumes under the term “language learner strategies” are the listening strategies in Vandergrift and Goh’s metacognitive model of listening. Although these “language learner strategies” do certainly play a role in language test situations and are an important part of the construct as discussed above, one could argue that they are not “test-taking strategies” as such. This is also acknowledged by Cohen, who argues that research in this area “can help us [...] to more rigorously distinguish language learner strategies on the one hand from test-taking (test management and test wiseness) strategies on the other” (A. D. Cohen, 2006, p. 325).

The second type of test-taking strategies in Cohen’s framework are test-management strategies. According to Cohen, as opposed to language learner strategies, test-management strategies are only relevant in test situations and are applied by test takers to deal with the specific demands of the test tasks. For example, Cohen lists the

strategy of “selecting options through the elimination of other options as unreasonable based on paragraph/overall passage meaning” (A. D. Cohen, 2006, p. 311). As this example shows, however, test-management strategies are also reliant on, or part of, language comprehension (“[...] based on paragraph/overall passage meaning”). To further illustrate this, the following table juxtaposes a list of test-management strategies for reading comprehension by Cohen with the listening strategies by Vandergrift and Goh discussed in the last section:

Table 1: Test-management strategies as defined by Cohen versus listening strategies as defined by Vandergrift and Goh

Examples of test-management strategies as defined by Cohen (2011, p. 307)	Vandergrift and Goh's (2012) listening strategies
Read the passage first and make a mental note of where different kinds of information are located.	Reorganizing (in double play)
Return to the passage to look for or confirm an answer rather than relying solely on memory of what was in the text.	Monitoring (in double play)
Read the questions first so that the reading of the text is directed at finding answers to those questions.	Planning
Read the questions a second time to make sure that their meaning is clear.	Planning
Try to produce your own answer to the question before you look at the options that are provided in the test.	Prediction
Make an educated guess – e.g., use background knowledge or extra-textual knowledge in making the guess.	Elaboration
Be ready to change the responses to any given item as appropriate – e.g., in the case where new clues are discovered in, say, another item.	Monitoring

As illustrated in the table, test-management strategies can be seen as being part of metacognitive strategies, in this case listening strategies as defined by Vandergrift and Goh, however, they are also informed by the test questions. This ambiguity is also indirectly acknowledged by Cohen, who refers to the strategies in Table 1 as “test-management strategies” (A. D. Cohen, 2011, p. 306) and then lists them under the heading of “[t]est-taking strategies which rely primarily on language use strategies” (A. D. Cohen, 2011, p. 307). Thus, it could be argued that test-management strategies, although specific to the test situation and reliant on the test questions, are dependent on language comprehension and therefore to a certain extent part of the construct. Until listening comprehension can be measured directly without the intermediary of the test paper or test questions, for example through sophisticated neuroimaging techniques

which might be available in the future, test-management strategies will be part of any test, simply because it is a test and not a real world situation. It could even be argued that test results would be less valid if test takers did not have any knowledge of test-management strategies. For example, if test takers are unfamiliar with a certain test format and therefore inhibited in their application of test-management strategies, the test results would not be an accurate reflection of the test takers' true knowledge. Still, test developers should make sure that the reliance on strategies which are informed by the test questions is kept to a minimum.

The last category in Cohen's model are test-wiseness strategies. Cohen defines these as "strategies for using knowledge of test formats and other peripheral information to answer test items without going through the expected linguistic and cognitive processes" (A. D. Cohen, 2006, p. 308). As testified by Field (2012, pp. 430–432), test takers frequently make use of test-wiseness strategies during listening tests. Test-wiseness strategies are clearly separated from language learner and test-management strategies as they are "test-dependent but language-independent" (Doe & Fox, 2011, p. 31). They therefore pose a greater threat to the validity of testing instruments as test-management strategies, because the latter are also dependent on the listening text and not solely the test questions. The following list, taken from Cohen, are typical examples of test-wiseness strategies:

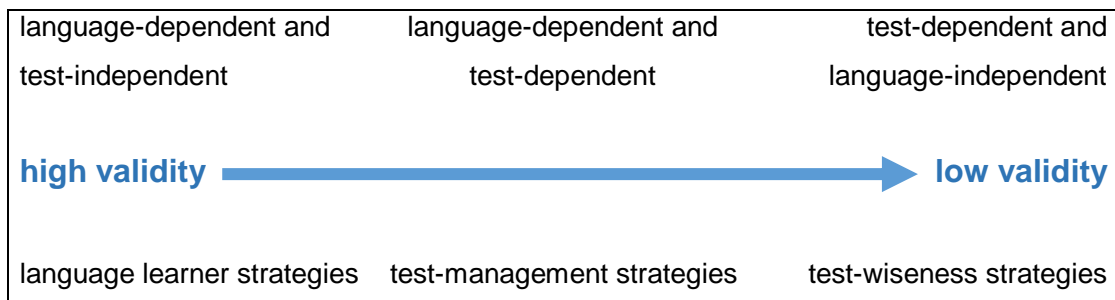
- Use the process of elimination – i.e., select a choice not because you are sure that it is the correct answer, but because the other choices don't seem reasonable, because they seem similar or overlapping, or because their meaning is not clear to you.
- Look for an option that seems to deviate from the others, is special, is different, or conspicuous.
- Select a choice that is longer/shorter than the others.
- Take advantage of clues appearing in other items in order to answer the item under consideration.
- Take into consideration the position of the option among the choices (a, b, c, or d).
- Select the option because it appears to have a word or phrase from the passage in it – possibly a key word.
- Select the option because it has a word or phrase that also appears in the question.

- Postpone dealing with an item or selecting a given option until later.
- Estimate the time needed for completing the items and don't spend too much time on any given item.

(A. D. Cohen, 2011, p. 307)

In summary, Cohen's three different types of test-taking strategies can be placed on a validity continuum, as illustrated in Figure 3. Test tasks targeting only language learner strategies, such as the listening strategies proposed by Vandergrift and Goh (see last section), have high validity. However, as any test is an artificial situation, test-management strategies will unavoidably play a role, because test takers cannot complete a test without them. For example, in multiple-choice tasks test takers need to choose an answer based on their understanding (a test-management strategy), but this does not necessarily jeopardise the validity of testing instruments, as language comprehension is still crucial to find the correct answer. On the other hand, when the use of strategies that are solely test-dependent but language-independent lead to correct answers (through test-wiseness), the validity of test scores is threatened.

Figure 3: Cohen's definitions of test-taking strategies on a validity continuum (based on A. D. Cohen, 2011; and Doe & Fox, 2011)



2.4.3.1. Potential effects of double play on the use of test-taking strategies

Double play could influence the use of all three types of test-taking strategies discussed above. The potential effects of double play on the use of language learner strategies in the context of listening were addressed in Section 2.4.2.1. The discussion here will focus on the potential impact of double play on test-management and test-wiseness strategies.

Test-management strategies, due to their reliance on language comprehension, could be affected in similar ways by double play as cognitive listening processes (see

Section 2.4.1.5) and listening strategies (see Section 2.4.2.1 and Table 1). In addition, as test-management strategies are also informed by the test questions, double play could influence their use. Test takers might rely more on the test question in single play situations, simply because they have fewer opportunities to apply “pure” listening strategies to find the correct answer. On the other hand, if test takers struggle to answer a question, they also have more opportunities to use the test questions in order to find a correct answer in a double play situation. Research has shown that test takers frequently use the test questions to inform their answers (see for example Field, 2012; Sherman, 1997), however, it has not been explored in detail how double play effects this.

Similarly, it is not yet clear how the use of test-wiseness is influenced by double play in listening assessment. It seems reasonable to hypothesise that tests utilizing single play might be more susceptible to test-wiseness. That is because in a single play condition, test-takers may be more likely to miss relevant information and therefore might rely more on guessing and other related strategies which are test-dependent but language-independent. It has not been established in detail how double play impacts the use of test-wiseness.

2.4.4. Anxiety

Apart from test-taking strategies², a potentially crucial variable threatening the validity of L2 listening tests is anxiety. Anxiety in relation to learning a foreign language has been defined as “the feeling of tension and apprehension specifically associated with second-language contexts” (MacIntyre & Gardner, 1994, p. 284). One form of language learning anxiety is L2 listening anxiety, which concerns negative feelings related to listening in a foreign language due to the unique features of spoken language, such as the need for real-time comprehension or lack of clarity (Vogely, 1998). L2 Listening anxiety has been shown to be empirically distinguishable from general language learning anxiety, however the two also share common characteristics and are thus regarded as different, but overlapping, constructs (Elkhafaifi, 2005; Kimura, 2008).

The detrimental effects of language learning anxiety are well documented (for a review of relevant literature see Horwitz, 2010), but research has been relatively sparse

² Based on the discussion in Section 2.4.3, henceforth the term “test-taking strategies” refers to test-management and test-wiseness strategies as defined by A. D. Cohen (2006, 2011), but not to language learner strategies.

in terms of how anxiety effects the different language skills. Still, a number of studies have shown that speaking, writing and reading in a foreign language can all be impeded by anxiety. Although research has been sparse in relation to listening, most studies agree that L2 listening anxiety negatively influences listening comprehension (Elkhafaifi, 2005; Golchi, 2012; Kim, 2000; Révész & Brunfaut, 2013). In all of these studies participants filled in a listening anxiety questionnaire after completing a listening test or a listening class and correlational analysis were conducted to investigate how test performance related to anxiety levels. Despite the use of a variety of task types (Elkhafaifi did not mention which tasks he used to assess listening performance) and different types of questionnaires, all of the studies found that listening performance correlated negatively with listening anxiety; that is, less anxious candidates in general performed better than more anxious candidates.

Another specific form of language learning anxiety pertains to the test-taking process itself. In contrast to listening anxiety, which relates to negative thoughts about listening in a foreign language, test-taking anxiety consists of “individuals’ cognitive reactions to evaluative situations, or internal dialogue regarding evaluative situations, in the times prior to, during, and after evaluative tasks” (Cassady & Johnson, 2002, p. 272). These reactions often include worries about ones performance compared to peers, concerns about possible failure, or lack of self-esteem. It has been established through a large body of research that high levels of test-taking anxiety are generally related to a decline in test performance (Hembree, 1988). Winke and Lim (2014), for example, used a questionnaire to investigate how test-taking anxiety relates to L2 listening performance. They found that test-taking anxiety negatively impacts test scores, thereby confirming findings described in Hembree (1988). In addition to their questionnaire analysis, Winke and Lim (2014) also compared eye-tracking metrics of low-anxiety candidates with those of high-anxiety candidates. The authors found that more test-anxious candidates spent significantly more time on reading the questions and took significantly longer to process the answer options, which suggests that less anxious candidates have more time to focus on the listening text.

In summary, although L2 listening anxiety and test-taking anxiety are different constructs in terms of the specific situations that elicit them, research has shown that both seem to manifest themselves by impeding language comprehension and they are also correlated with a decline in test performance. However, it has not been investigated in detail how double play could influence either of these constructs.

2.4.4.1. Potential effects of double play on anxiety

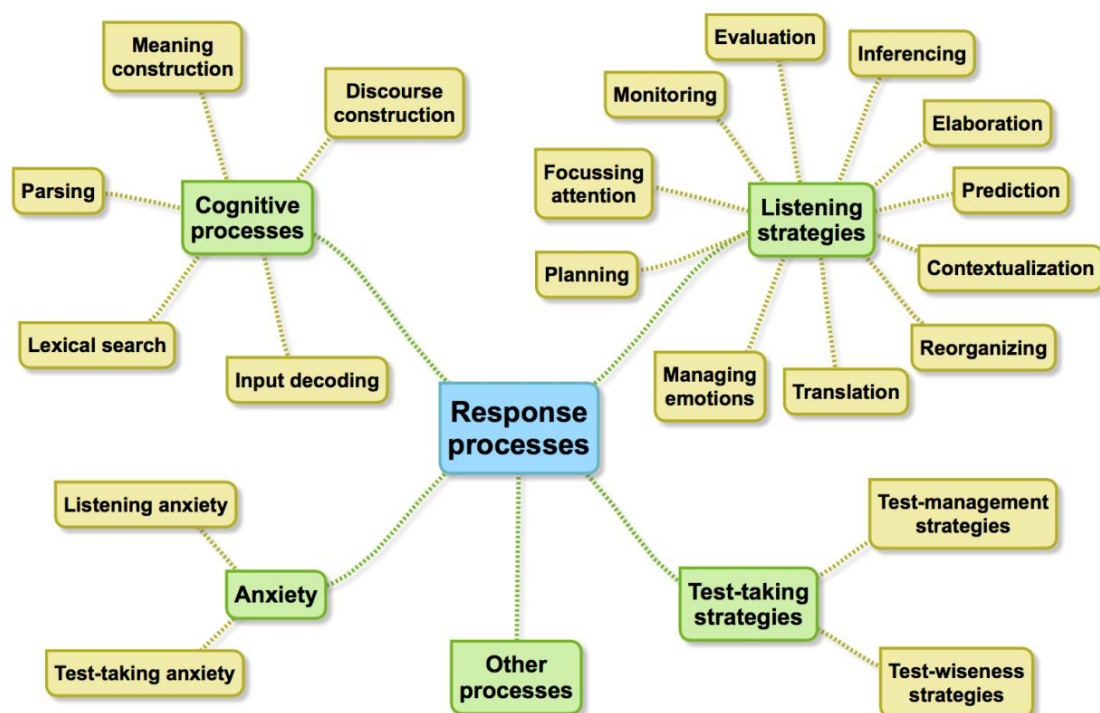
It seems reasonable to hypothesise that double play could lower candidates' anxiety levels. If students know from the outset that they will hear the listening text a second time, they may feel less intimidated by having to listen in a foreign language. They may also be less worried and less stressed about taking the test and consequently more confident in their abilities.

Field's (2015) research suggests that double play might indeed reduce anxiety – however, due to the research design Field was not able to compare anxiety levels of candidates who experienced single play with candidates who experienced double play. In his stimulated recall study (see Section 2.5.4), out of the 19 candidates who reported on anxiety, 15 indicated that they felt less anxious during the second play in a double play condition as compared to the first play. Although this finding is interesting, the extent to which double play influences anxiety levels as compared to single play remains to be explored.

2.4.5. Summary of the theoretical framework

In summary, the theoretical framework of response processes in listening assessment consists of four main dimensions: cognitive processes and listening strategies (response processes of interest), and test-taking strategies and anxiety (irrelevant response processes). All of these are divided into various sub-dimensions, as illustrated in Figure 4. These dimensions provide the foundational validation framework for the thesis.

Figure 4: Summary of the theoretical framework of response processes in listening assessment



2.5. Previous research on double play in listening assessment

A number of peer-reviewed empirical studies have investigated the effects of double play in listening assessment. Most of these investigations looked at double play as a secondary part in a larger study, while only a few focused exclusively on double play. The focus of most studies was whether or not repeating the listening text had an effect on test takers' scores, or whether it changed item properties such as discrimination indexes, and one study investigated the effects of double play on item reading and answer changing. Given the relatively small literature on this topic, in the following sections the results of the individual studies will be outlined in detail before a synthesis of the findings is presented. Particular attention will then be given to the study by Field (2015), as a literature review revealed it to be the only large-scale study investigating response processes in relation to double play to date.

2.5.1. Studies looking at double play as part of a larger study

The earliest published study which addressed double play appears to be Lund's (1991) investigation of listening and reading abilities of 60 beginner and intermediate college

students of German. For reasons of space and relevance, only the results on the effect of double play on listening performance will be discussed here. Lund compared the two groups of learners by means of the free recall method and found that participants experienced “significant benefits” (Lund, 1991, p. 201), in terms of the both quantity and accuracy of recalled lexical items and propositions, from the second play of the text. Intermediate learners benefitted more than beginner learners. Lund admitted that these benefits can to a certain extent be attributed to the research method used, i.e. the fact that participants produced a free recall protocol after the first listening may have helped them remember lexical items and propositions, thereby aiding comprehension during the second listening. However, he argued that the second listening is crucial for achieving increased understanding (Lund, 1991, p. 201).

Berne (1995) provides further evidence for the performance benefit of double play. In this study, the main focus of which was to investigate the effects of varying pre-listening activities on listening performance, double play was again investigated as a secondary treatment. Berne randomly split 62 English-speaking learners of Spanish into three experimental groups, two of which performed a pre-listening activity (either previewing the questions or performing a vocabulary preview activity). After the pre-listening activities, the participants watched a video-taped lecture and completed a multiple-choice task, followed by a free written recall. Following the written recall, the recording was played again and the participants re-took the multiple-choice test. The results after the first listening indicated that participants benefited more from previewing the questions than from studying passage related vocabulary. However, after the second listening all three groups performed significantly better than after the first listening, regardless of the type of pre-listening activity. The author thus concluded that “the most effective means of improving listening comprehension performance is through additional exposure to the passage” (Berne, 1995, p. 326).

A different approach to investigating double play was taken by Sherman (1997), who specifically looked at the effects of different types of question preview on 178 intermediate-level undergraduate students’ performance. Four groups of test takers performed four different listening tests under four different conditions. Group A previewed the questions first and then listened to the recording twice. Group B heard a different passage twice before seeing the questions. Group C listened to the third recording once, viewed the questions and listened a second time (called the “sandwich” method). Group D did not answer questions but performed a free written recall after

listening to the fourth passage. All of the groups completed a questionnaire immediately after the tests and produced another free written recall of their respective passages one week after the initial hearings. The test results indicate that in terms of overall score, group C (the “sandwich” group) significantly outperformed the other groups. In addition, questionnaire results suggest that test takers preferred the sandwich version over the other versions and found it less effortful, less distractive, and less tense. Although test results of group A (question preview before the first hearing) and group B (access to questions only after the two hearings) were very similar, test takers by far preferred version A to version B. Sherman’s research design, however, raises several questions. It is not clear from the design whether the performance benefits can be attributed to the listening condition, the listening passage, or the specific group completing the task. The results of the written recalls one week after the tests were inconclusive, as no significant differences between the four groups were found. However, this last finding may be attributable to memory effects, as after one week memory probably decayed for all groups so that only the bare information was recallable.

A. C.-S. Chang and Read (2006) provide additional evidence for increased test scores through double play. They compared the listening test performance of four groups of 40 Taiwanese learners of English at a low intermediate proficiency level. Each of the groups answered multiple-choice questions on two listening passages under different conditions. The first group heard the recording twice. After the first hearing, test takers were handed the questions for preview before the second hearing. The second group read a topic preparation text prior to the listening, which summarised the content of the two listening texts, followed by a 25-minute discussion of the topics. They previewed the questions before hearing the recording once. The third group studied vocabulary relevant for the listening passages for 25 minutes, followed by 25 minutes of instruction on pronunciation and in-context-usage of the vocabulary. Equal to the second group, they previewed the questions before listening to the passage once. The fourth group only previewed the questions, but did not receive any other support, and therefore served as the control group. Statistical analyses of test scores revealed that the topic preparation group received the highest scores, followed by the double play group, the question preview group and the vocabulary instruction group. The researchers concluded that the high scores of the topic preview group showed “that having a relevant content schema means that students are less dependent on vocabulary

knowledge in achieving adequate comprehension” (Chang & Read, 2006, p. 392). Regarding double play, the results of Chang and Read’s investigation showed that students preferred it and achieved higher scores than the control group. Higher proficiency students seemed to benefit more than lower proficiency students, a results which confirms Lund’s (1991) findings.

Contrary to these findings, Henning (1991) reports no benefits of double play. Among other variations in test conditions, he looked into the effect of double play on 120 test takers’ performance on three listening tests based on 144 TOEFL test items, which were partially adapted for the purpose of the study. The other conditions investigated were length of the listening passage, length of response options, and level of cognitive processing required. Henning found that double play tended to decrease item difficulty, but this did not reach statistical significance when mean difficulty was compared with single play. His results also showed that double play did not have a positive effect on item discrimination, on item response validity (as indicated by fit statistics), or on format construct validity (as indicated by a correlation matrix). Based on these findings, Henning implied that there is no need to repeat the listening texts in the TOEFL test, a practice which still holds today.

Similar findings are reported by Brindley and Slatyer (2002). In a study on exploring listening task difficulty, they examined the effect of double play among other factors such as speech rate, text type, input source, and item format. The researchers used three different tasks, two of which were changed according the varying task characteristics and task conditions under scrutiny, and the third was used as a control task. All of the 284 participants completed three tasks. The control task was the same for all participants, while the different versions of the two tasks with varying characteristics and conditions were randomly assigned. The authors found that speed of delivery and item format affected task difficulty, but double play did not seem to have an effect. Overall, however, the authors concluded that “the complexities of the interactions between task characteristics, item characteristics and candidate responses [...] suggest that simply adjusting one task-level variable will not automatically make the task easier or more difficult” (Brindley & Slatyer, 2002, p. 290). Brindley and Slatyer suggested conducting introspective studies in order to investigate these complex issues in more detail.

2.5.2. Studies focussing exclusively on double play

Apart from the studies outlined above, which have looked at double play as a secondary treatment in a larger study, a limited number of studies have focussed exclusively on double play. The first one was published by Iimura (2007), who explored the effects of double play in relation to different question types. In this study, 165 Japanese senior high-school learners of English were divided into three proficiency groups, based on the students' results on the listening section of the 3rd grade level of the Society for Testing English Proficiency (STEP) test. In the main study, all of the participants listened to 10 short passages (18-20 sec) from the pre-2nd grade STEP test. Following each passage, the test takers had to answer one question targeting main ideas in the passage (global question) and one question targeting local details (local question), before listening to the passage again and answering the questions a second time. Results showed that all participants scored significantly higher after the second listening, both for global and local questions, than after the first listening. Learners with higher proficiency benefited more from the second listening than those with lower proficiency, which confirms Lund's (1991) and A. C.-S. Chang and Read's (2006) results. Regarding the different question types, the author did not find significant differences between proficiency levels, nor between the first and second hearing of the passage. However this last finding might be explained by a lack of precision in the instrumentation used in the study. A closer look at the questions reveals that the difference between global questions and local questions is not always clear. An example is shown below of an excerpt from the listening test used in the study, along with the associated global and local questions:

Listening text: *When Sara was a child, she loved reading books. In high schools [sic] she began writing stories as a hobby. When she was 16, she entered a short-story contest and won first prize. Now she makes a living by writing books for children. And she still loves reading.*

Global question: *What does Sara do now?*

Local question: *When did Sara win first prize?*

As shown in the excerpt, the difference between the two question types is not clear, as both questions seem to target local information. The author acknowledges this limitation and also points out that for such short passages it might be difficult to target

processing at a global level. Thus, although the study confirmed that double play increased test takers' performance, the findings do not offer conclusive insights on effects of double play on different question types.

Another investigation of double-play which supported a double play advantage was conducted by Sakai (2009). Sakai investigated the performance of students of different proficiency levels when listening to a passage twice. Thirty-six Japanese-speaking university students of English were divided into two proficiency groups, based on their results on the listening sections of three forms of the Michigan English Placement Test. They then listened to two listening passages lasting 27-29 seconds. After each passage was played, the participants were given three minutes to write down everything they could remember from the passage. The procedure was repeated a second time. The written recall protocols were analysed by counting the number of idea units retained and by looking more closely at recall units which were not part of the original passage but fitted the context of the listening text (referred to as "idiosyncratic" recall units). The results show that both proficiency groups scored significantly higher on the second listening, and also improved in terms of idiosyncratic recall units produced. Sakai concluded that his findings do not lend support to the notion that double play has differential effects on students of different proficiency levels, but that it is beneficial regardless of level of proficiency of test takers.

In a more recent study, Ruhm, Leitner-Jones, Kulmhofer, Kiefer, Mlakar, and Itzlinger-Bruneforth (2016) investigated the effects of double play on test outcomes of 1,266 14-year old L2 learners of English at A2/low B1 level of the CEFR. The participants were divided into two groups and each completed 20 multiple-choice items in a single play condition (group 1, N=443) or a double play condition (group 2, N=823). The items were based on either a "short" stimulus lasting up to 60 seconds and a "long" stimulus of two minutes on average. It is not clear how the items were developed, i.e. whether they were field trialled or standard set. Ruhm et al.'s results show yet again that in general item difficulty decreases when the listening text is played twice. Their findings with regards to stimulus length suggest that for long stimuli item difficulty decreases more than for medium-length stimuli, although it is not clear what the authors mean by "medium-length" stimuli. In addition, Ruhm et al. found that items with low difficulty in single play benefit less from a second play in terms of decreased item difficulty than items with high difficulty in single play.

Instead of investigating the effects of double play on item difficulty, Aryadoust (2019) used eye-tracking to study the extent to which test takers focus on item prompts and answer options and change their answers in computer-delivered listening tasks featuring double play. In his study, 28 secondary school students completed six multiple-choice items and six matching items while their eye-movements were recorded on an eye-tracker. Aryadoust then compared the amount of attention, as measured by eye-movement metrics such as total average fixation duration, average fixation count, average visit duration on specified areas of interest, and average visit count, between the pre-listening stage of the first play, the while-listening stage of the first play, the pre-listening stage of the second play, and the while-listening stage of the second play. The author found that students focussed significantly longer on the items in the while-listening stages than the pre-listening stages, particularly during the second play. In addition, attention (as measured by eye-movement metrics) to the questions and answer options was greater for the matching items than the multiple-choice items. In terms of changing answers, in about two thirds of all cases candidates changed their answers from incorrect to correct. Based on these findings, Aryadoust argues that listening tests where candidates need to simultaneously read and listen are not “an authentic representation of real-life listening processes” and he suggests to “[eliminate] the multitasking requirement to maximise the authenticity of such tests” (Aryadoust, 2019, p. 21). However, Aryadoust does not offer any recommendations as to how listening comprehension could be tested without the reading of test questions. In addition, the underlying assumption of Aryadoust’s study that eye-movement patterns during listening tests are indicative of attention to reading can be challenged, as will be discussed in detail in Section 7.5.

2.5.3. Summary of the findings on studies of double play

In sum, the great majority of the investigations on double play in listening assessment found that it aided comprehension and increased test takers’ scores (Berne, 1995; Chang & Read, 2006; Iimura, 2007; Lund, 1991; Ruhm et al., 2016; Sakai, 2009). More proficient test takers seem to benefit more from double play as less proficient test takers (Chang & Read, 2006; Iimura, 2007; Lund, 1991), although one study found no effects in terms of proficiency (Sakai, 2009). A smaller number of studies reported that students did not benefit from double play as much as expected (Brindley & Slatyer, 2002;

Henning, 1991). In addition, the studies looking at different task types in relation to test scores seem to agree that double play might not be mediated by task type. Finally, a recent study by Aryadoust (2019) found that students focussed longer on the test questions during the second play compared to the first as indicated by eye-movement metrics, however it is not clear whether the increased fixation and visit durations are indicative of higher reading load, more intense listening, or a combination of these and other factors (see also discussion in Section 7.5).

The effect of double play on test outcomes and candidates' eye-movement patterns is worth addressing, but it could be argued that the more relevant research question for making decisions about language test design is whether repeated input has any implications for the coverage of the listening construct. None of the studies outlined above investigated this in detail. This can be explained to a certain extent by the fact that some of the research was framed in pedagogical terms and was not focussing on testing (Berne, 1995; Lund, 1991). Also, as Vandergrift and Goh (2012, pp. 5–6) point out, most research in teaching listening has mainly focussed on the product of the listening process, rather than the listening process itself. This might also explain why the great majority of studies in the field of language testing (Boroughs, 2003; Brindley & Slatyer, 2002; Chang & Read, 2006; Henning, 1991; Ruhm et al., 2016; Sakai, 2009; Sherman, 1997) did not investigate the effects of double play on the listening construct, but only its effects on test difficulty, or on test taker preferences. Ideally, however, the construct should be the main focus in studies addressing this topic. This, as Buck (2001, p. 1) stresses, “is **construct validity**, and ensuring that the right construct is being measured is the central issue in all assessment” (bold in original).

In particular, it has not been fully established in what ways multiple exposures to the listening text influences test takers' response processes. As Buck (2001, p. 171) observes, “hearing and processing a text a second time may utilise different comprehension skills from the first time – we really do not know”. As mentioned above, the majority of studies which have addressed double play thus far are not helpful in this respect, because “the rather simplistic notion of ‘difficulty’ as reflected in item difficulty statistics is of limited usefulness in understanding what happens when an individual candidate interacts with an individual item” (Brindley & Slatyer, 2002, p. 390). Bachman (1990) put it similarly when he wrote that a “[...] critical limitation to correlational and experimental approaches to construct validation [...] is that these

examine only the products of the test taking process, the test scores, and provide no means for investigating the processes of test taking themselves” (p. 269).

Instead of solely focussing on the product (i.e. the test score), researchers need to also investigate the test-taking process from the perspective of the test taker to more fully understand the complex nature of assessment. As pointed out by Hubley and Zumbo, “[i]dentifying and understanding the mechanisms underlying how different respondents interact with, and respond to, test items and tasks is essential to understanding score meaning and test score variation” (Hubley & Zumbo, 2017, p. 8). Similarly, Weir (2005) asserted that investigating the processes underlying test-taking is crucial. Field (2015) was the first researcher to conduct a full-scale study on the effects of double play on test takers’ response processes and, given the relevance of his study to the current research, his findings will be outlined in detail below.

2.5.4. Field 2015

Field’s study builds on the findings of two small scale investigations (Buck, 1990; Field, 2009), which indicated that test takers may indeed utilise different response processes in single play and double play listening tasks. The aim of Field’s study was “to examine whether the two different listening conditions materially affect candidate behaviour and candidate scores” (Field, 2015, p. 5), in order to provide guidance for the development of the Aptis listening test. In his investigation, Field first collected quantitative data of 73 participants taking two IELTS listening tasks (multiple-choice and gap fill). Before data collection, participants were told that they would hear the listening text only once, but after the first play they were informed that they could listen again. Field compared the scores between the first and second play and analysed in what ways participants changed their answers. The findings suggest that although there was a general increase in test scores after the second play across all proficiency levels, there were strong individual differences. In addition, the gap-filling task benefitted significantly more from the second play in terms of increased test scores than the multiple-choice task, as many test takers found it hard to answer gap-filling questions after only one play. However, it is not clear whether this was due to the nature of the task format, what was being targeted by the items, or a combination of the two.

In the second part of the investigation, Field analysed stimulated recall protocols of 37 participants while they were solving the same two IELTS listening tasks.

Interestingly, for this part of the study, participants were told from the outset that they would hear the text twice. During both plays, the recording was stopped after several questions and participants were asked how they had arrived at their answers (during the first play) and how they had made use of the second play (during the second play). Field reports that during both the first and second play the majority of participants heavily relied on word level decoding. However, he stresses that overall their behaviour “differed markedly” between the two conditions, as only during the second play many participants made use of higher-order listening processes to understand the overall meaning of the passage. Participants also reported lower levels of listening anxiety and greater familiarity with the content during the second play.

Field’s study is important as it is the first to focus on test takers’ response processes in relation to double play, but there are still questions which remain unanswered. For example, in his qualitative study Field did not compare the response processes of test takers who know that they are only allowed one hearing and test takers who know that they are allowed to listen twice. Although he initially planned such a comparison he ended up focussing only on the double play condition, arguing that “it proved hard to conclusively identify marked differences of behaviour during a single play as compared with the first hearing of a double play” (Field, 2015, p. 12). However, a more separated look at single play versus double play would be insightful. Especially in terms of test-taking strategies and anxiety there might be differences in processing between candidates who know from the outset that they only get one chance at understanding and candidates who know that they will get a second chance after the first listening. Also, test-takers might listen more intensively and thus differently in a single play condition than in the first hearing of a double play condition.

Another avenue to explore to potentially gain a fuller understanding of candidate’s processes concerns the chosen recall methodology. Field used the participants’ answers as stimulus in his recalls and in some instances appears to assist the participants in their answer choice on the first listening, as the following example of a recall during a first hearing shows:

S: yeah ++ and so I choose A + I think maybe A is much more important

R: + um ok + but would you like to read A? + have + have a look at A

S: ++ look at what?

R: well

S: uh huh?
R: read A again
S: oh read A + ok
R: to be sure that it's right
S: (mutters)
R: ok?
S: oh
R: we'll stay with A for the moment yeah?
S: uh huh
R: um + but you're going to have a chance to listen to it again + ok?
S: ok (Field, 2015, p. 52)

In this particular instance the recall quite likely influenced the participants' processing during the second play, as they may have paid more attention to answer A due to the hint the researcher gave that it may not be the correct answer.

Another excerpt of a stimulated recall during the first play is presented below. In this instance Field assumed that the participant was guessing, but the participant did not seem sure that they were. They might just have not fully focussed on the item yet, as they knew that they would hear the recording again:

R: I presume you heard Europe as well did you?
S: yeah
R: so it was a little bit of a guess was it?
S: yeah maybe + it's about guess yeah (Field, 2015, p. 52)

More neutral types of prompts within a stimulated recall procedure would be likely to help gain a clearer understanding of test takers' processing. For example, recalls could be conducted without using the answers to the items as stimulus, thus avoiding the constraint the items put on the recall procedure. The advantage of such an approach would be that researchers might get a more general overview of participants' processing, by allowing room to observe more in terms of what is being understood, and how listeners are achieving this understanding. At the same time, response processes in understanding the items would likely emerge from the data as well. Following this procedure would also avoid the problem of 'forcing' participants to talk

about what they were doing on a particular item during the first play, even if they had not actually fully focused on the item yet. So the initial probe question would be “Can you tell me what you were thinking during that section?”, rather than “Can you tell me what you were thinking when you answered that question?”.

In addition, by interrupting the audio during the first and second play and conducting stimulated recalls, Field might have altered test takers’ natural processing, an effect which is referred to as the “reactivity” problem (Gass & Mackey, 2000; Russo, Johnson, & Stephens, 1989). In particular, test takers’ processing during the second play might have been influenced by the stimulated recalls during the first play. Stimulated recalls conducted only after task completion could help avoid such reactivity effects, although a potential disadvantage could be the greater time delay and difficulty to distinguish between first and second play processing.

Field used English in the stimulated recall questions and participants had to report in English, which was the target language and participants’ L2. Although he asserts that “[t]he precise wording of the questions was adjusted somewhat to fit the participant’s level of English and powers of self-expression” (Field, 2015, p. 20), this practice might still have had an impact on the results. Leaving participants the choice between L1 and L2 (or mixed) for the reports might yield more detailed and less restricted recalls, and thus might help gain a deeper understanding of participants’ processing. As Bowles (2010) points out, if participants are asked to use their L2 for the recalls, they “might [...] be unable to communicate some of their thoughts as effectively as they could in the L1” (Bowles, 2010, p. 115).

One important finding of Field’s research was that participants’ scores on the gap-fill task improved markedly more after the second play than scores on the multiple-choice task. This was because participants often struggled to answer gap-fill questions during the first play, as they had to listen and simultaneously write down the answer to the question. It remains to be explored in what ways variations in task types are affected by single versus double play. For example, it would be interesting to investigate whether other open format tasks where test takers need to answer questions or complete gaps at the end of sentences are affected in the same way, or whether the number of options in multiple-choice tasks impacts test takers’ processing with regards to single and double play.

3. Methodology

It was identified in the literature review that 1) the main research gap in relation to double play in L2 listening assessment concerns the lack of focus on test takers' response processes (cognitive processes, listening strategies, test-taking strategies, and anxiety) and 2) existing studies, although insightful, could be improved and expanded upon in terms of research design, research methodology, and task types used. Based on this reasoning, in this chapter of the thesis relevant research questions will be formulated first (Section 3.1), followed by a discussion of methods best suited to answer these questions (Sections 3.2 and 3.3). Next, the general research design will be described in Section 3.4, including the research context, the pilot studies conducted, and the tasks used in the research. The following sections then outline the specific research design of Study 1 (Section 3.5) and Study 2 (Section 3.6), including the participants, the additional materials used in the research, as well as the analysis procedures.

3.1. Research questions

In order to compare the findings with the main share of previous research in this area, the following question will first be investigated:

1. What are the differences in item and task statistics between listening tasks completed in single play and double play?
 - a. Is task type a factor?

The following research questions then relate to test takers' response processes. As outlined in Section 2.4, the theoretical validation framework of the thesis consists of four main response processes: cognitive processes and metacognitive strategies (the construct of interest), and test-taking strategies and anxiety (construct-irrelevant factors). The thesis aims to identify how double play affects these four dimensions. Thus, based on the identified research aims the following research questions will also be addressed:

2. What are the differences in test takers' cognitive processing between listening tasks completed in single play and double play?
 - a. Is task type a factor?
3. What are the differences in test takers' use of listening strategies between listening tasks completed in single play and double play?
 - a. Is task type a factor?
4. What are the differences in test takers' use of test-taking strategies between listening tasks completed in single play and double play?
 - a. Is task type a factor?
5. What are the differences in test takers' anxiety levels between listening tasks completed in single play and double play?
 - a. Is task type a factor?

Investigating these questions will inform listening researchers not only on the effects of double play on the validity of listening assessment instruments, but should also be insightful on a more general level. By employing the methods outlined in the following, the study will look into cognitive processes, metacognitive strategies, and anxiety levels of L2 listeners and the results will thereby help deepen the understanding of the listening construct. As outlined in the introduction, this is important since “from a primarily cognitive perspective, the processes involved in second language listening are perhaps the least well described and analysed in the currently available literature on language assessment” (Taylor & Geranpayeh, 2013a, p. 326). However, the research also goes beyond assessment as it will help us to understand how repeated listening affects comprehension processes more generally.

3.2. Methods to analyse test-related data

Broadly speaking, there are two methods to statistically explore test-related data: Classical Test Theory (CTT) and Item Response Theory (IRT, sometimes also referred to as Modern Test Theory). In this section the two methods will be outlined briefly.

CTT can be used, among other applications, to explore item and task difficulty (i.e. the number of correct responses for each item), item discrimination (item-total correlation), and test reliability (e.g. Cronbach's Alpha). In simple terms, CTT puts a persons' assumed true score on a test in relation to the observed score and the associated

measurement error (Magno, 2009). An advantage of CTT is that it can be used with common statistical programs such as SPSS, which enables practitioners to calculate basic test indices in a familiar environment without the need for more advanced programming skills (Green, 2013, p. xiii). One major drawback of CTT, however, is its dependence on the test taker population, particularly with small sample sizes (Hambelton, 2000; Magno, 2009). Thus, unless the test population is truly representative of the target population, CTT results are often not generalizable.

In contrast to CTT, which is largely dependent on correlational analyses, IRT is based on probability theory and considers both person ability and item difficulty in calculating the chance of a person getting a particular item correct (Kaplan & Saccuzo, 1997). As such, IRT has stronger underlying assumptions than CTT (Magno, 2009) and “makes it possible to estimate sample-free item difficulty and item-free person ability” (Green, 2013, p. xiii). In other words, IRT enables researchers and practitioners to directly compare person ability and item difficulty across multiple scenarios. Some disadvantages of IRT, on the other hand, are that larger sample sizes are generally required (a minimum of 200 participants is often suggested in the literature) and that associated software is usually less user-friendly and requires some basic knowledge of programming language.

3.3. Methods to study test taker response processes in listening assessment

A persistent challenge for investigating test taker cognition is the choice of research method. Language assessment researchers have a number of methodological options, none of which is without disadvantages (Purpura, 2013, p. 19). In an overview of the topic, Gass and Mackey (2007) list the following techniques:

- Direct observation
- Carefully structured observation (such as eye-tracking sensors)
- Tracking behaviour (e.g. via computer using keyboard tracking programmes)
- Questions, either direct (e.g. interviews) or indirect (e.g. questionnaires)
- Obtaining retrospective information (e.g. in the form of diaries, or through retrospective recall procedures)

- Online commentary (e.g. learners speak their thoughts aloud while they are engaged in an activity)

(adapted from Gass & Mackey, 2007, pp. 45–46)

Clearly, due to the nature of the task, some of these methods are not applicable for researching cognitive processes and metacognitive strategies in listening assessment. For example, while observation techniques such as the use of computer tracking programmes might be helpful for investigating test takers' response processes in writing tests, they are less applicable for listening tests, as the main form of input is oral and not written. Similarly, participants cannot think aloud while listening to passages and answering test items, but have to recall their thoughts retrospectively (Goh, 2002, p. 189). For these reasons, listening cognition researchers are somewhat limited in their choice of methods, and in most situations have to rely on questionnaires and retrospective recall protocols. In addition, the use of eye-tracking has recently been shown to be useful for stimulating retrospective recalls, helping to mitigate memory effects (Brunfaut & McCray, 2015; Holzknicht et al., 2017). In the following, the use of questionnaires and verbal protocols to investigate response processes in listening assessment will be reviewed in more detail.

3.3.1. Questionnaires

Collecting data by means of questionnaires is one of the most popular methods in applied linguistics and in the social sciences more generally. This method has also been used extensively in research on test takers' response processes (Hubley & Zumbo, 2017, p. 4). As defined by Brown (2001), “[q]uestionnaires are any written instruments that present respondents with a series of questions or statements to which they are to react either by writing out their answers or selecting from among existing answers” (Brown, 2001, p. 6). According to Dörnyei and Taguchi (2009, pp. 5–6), questionnaires can elicit factual, behavioural, and attitudinal data from respondents. Factual questions target demographic and other background information. Behavioural questions are aimed at gathering data about respondents' past actions, for example their use of particular response processes such as listening strategies. Attitudinal questions, on the other hand, target respondents attitudes to a domain of interest.

As outlined in detail in Dörnyei and Taguchi (2009), researchers need to pay attention to a number of general factors when constructing a questionnaire. Questionnaires should not be too long, they should have an appropriate and attractive layout (see also Sanchez, 1992), sensitive topics should be avoided, and respondents should be reminded that their responses will be treated anonymously. In addition, it is essential that questionnaires contain clear instructions so respondents know exactly what they are asked to do. Ideally, questionnaires should be piloted before the main data collection in order to inspect the clarity of instructions, to detect problematic items (e.g. due to unclear wording), and to test the administration procedure.

The most commonly used item type in questionnaire research are closed-ended statements to which respondents indicate their level of agreement on a Likert scale, such as from “strongly agree” to “strongly disagree”. Two advantages of this item type are that responses do not need to be coded, which avoids introducing subjectivity into the data, and that they can easily be transformed into numerical data for quantitative analyses (Dörnyei & Taguchi, 2009, p. 26). In terms of number of Likert scale categories, researchers often recommend using an even number in order to avoid respondents’ tendency of opting for the neutral middle option (e.g. “neither agree nor disagree”), as it has been shown that middle options are often chosen by less motivated respondents (Krosnick, Judd, & Wittenbrink, 2005). In addition, middle categories can also be sensitive to cultural biases, for example they are more often chosen by Asian students compared to students from North America (Chen, Lee, & Stevenson, 1995).

In their comprehensive overview of the topic, Dörnyei and Taguchi (2009) also offer detailed guidelines with regards to the administration of questionnaires. First and foremost, researchers should ascertain that the participant sample is representative of the target population. In addition, in order to carry out subsequent statistical tests such as factor analyses, the sample size should be at least 100 people, but more are generally better to achieve meaningful results. It is also important to carefully plan the actual administration. Group-administration (e.g. in a language classroom) is often preferred to self-administration (e.g. as an online survey), as the procedure is more standardised and usually results in higher response rates (see also Wagner, 2012, p. 2).

When it comes to analysing questionnaire data, researchers have a number of options. For the commonly used Likert scales, statements measuring the same construct can be combined to multi-item scales by conducting a factor analysis (Hatch & Lazaraton, 1991). This procedure has the advantage that researchers can check whether

those items which were meant to tap into the same domain are actually measuring the same construct. In addition, multi-item scales reduce the number of variables in the analysis, which simplifies interpretation of the data (Dörnyei & Taguchi, 2009, p. 91). Once the items have been grouped into multi-item scales, it is important to check whether the measurements for each domain are reliable using statistical procedures. For the commonly used Cronbach's Alpha reliability index, values greater than .70 are desirable (Vogt, 2007). If the questionnaire study is based on a large enough sample size, inferential statistics can be used to inspect statistical significance in order to generalise findings onto a larger population.

Although questionnaires are very popular in the social sciences, they also have some drawbacks. One criticism relates to the fact that psychological domains such as the use of listening strategies are highly abstract and questionnaires can only measure them indirectly through a limited number of items (Wagner, 2012, p. 4). Other common criticism focus on the lack of control over the respondents' answers: researchers cannot be sure whether respondents are motivated enough to answer truthfully, and erroneous answers can usually not be corrected (Dörnyei & Taguchi, 2009, pp. 6–9). However, some of these limitations can be mitigated by taking care in constructing and administering questionnaires as outlined above.

3.3.2. Verbal protocols

Another common technique to investigate test takers' response processes is by means of verbal protocols (sometimes also referred to as verbal reports or think-aloud protocols). This method has “become intrinsically intertwined” with research on response processes as part of collecting validity evidence (Hubley & Zumbo, 2017, p. 3). Gass and Mackey (2000) define verbal protocols as “gathering data by asking individuals to vocalise what is going through their minds as they are solving a problem or performing a task” (2000, p. 13). A landmark theoretical framework for verbal protocols was developed by Ericsson and Simon (1987, 1993). Since then, this technique has been used in a wide range of fields, including L2 listening research.

There are different forms of verbal protocols. Ericsson and Simon (1987, 1993) differentiate between “concurrent” and “retrospective” verbal reports. In concurrent reports participants think aloud while they are engaged in the activity, whereas retrospective reports are generated some time after participants finished the activity. It

is generally agreed that concurrent reports tend to be more valid, as they are “less susceptible to influences from unwanted variables than are retrospective reports” (Green, 1998, p. 6). However, as outlined above, concurrent reports cannot be used for researching listening processes due to the nature of the task. In such cases, Ericsson and Simon suggest conducting retrospective reports immediately after the activity is finished (1987, pp. 40–41). The time frame between the activity and subsequent recall is crucial. According to Ericsson and Simon, “[...] due to the limited capacity of STM [short term memory], only the most recently heeded information is accessible directly. However, a portion of the contents of STM are fixated in LTM [long term memory] before being lost from STM, and this portion can, at later points in time, sometimes be retrieved from LTM” (1993, p. 11).

A different distinction between verbal reports is drawn by A. D. Cohen (1987, 1996, 2011), who discusses the method specifically in relation to L2 strategy research. He differentiates between self-report, self-observation and self-revelation (A. D. Cohen, 1996, p. 13). In self-reports, learners describe “what they can do” in the form of “generalized statements about learning behaviour”. Self-observation involves reporting “specific rather than generalized behaviour, either introspectively, i.e., within 20 seconds of the mental event, or retrospectively” (A. D. Cohen, 1996, p. 13). In self-observation learners thus not merely report but also analyse their thought processes. In contrast, self-revelation is described by Cohen as a “stream-of-consciousness disclosure of thought processes”, and is thereby closest to Ericsson and Simon’s definition of verbal reporting.

A specific form of verbal protocols, which has been described in detail by Gass and Mackey (2000, 2007; Mackey & Gass, 2005), is the stimulated recall method. In A. D. Cohen’s terms, stimulated recall can be described as a form of retrospective self-observation. In contrast to other retrospective verbal report techniques, stimulated recall is characterised by the use of a stimulus, the purpose of which is “to reactivate or refresh recollection of cognitive processes so that they can be accurately recalled or verbalized” (Gass & Mackey, 2000, p. 53). This stimulus can have different forms, and should be “some tangible (perhaps visual or aural) reminder of an event” (Gass & Mackey, 2000, p. 17). For investigating cognitive processes of listening test takers, the use of individual test items as stimuli might be helpful. This is also observed by Field, who argues that “the circumstance of a listening test support retrospection well in that the participant

has to provide a set of answers, which provide triggers to assist recall of the thought processes that led to them” (Field, 2012, p. 35).

In more recent research, participants’ eye-movements have also been used as a stimulus to initiate retrospection. Although eye-movement metrics alone are of limited usefulness for studying listening test takers’ response processes, as the influence of the listening text on test takers’ eye-movements cannot be untangled from their reading behaviour (Salverda, Brown, & Tanenhaus, 2011; Winke & Lim, 2014), using participants’ eye-movements as stimulus for verbal recalls has been shown to provide rich and potentially novel insights (Brunfaut & McCray, 2015; Holzknecht et al., 2017; McCray et al., 2012; McCray & Brunfaut, 2016; Winke & Lim, 2014). In these studies, participants saw a video of their eye-movements while they had been solving the items to help them remember their thought processes. The authors report that this procedure was unobtrusive and that the eye-movements served as a powerful stimulus to help participants recall their thought processes.

As with every research method, there are some constraints associated with verbal protocols. Gass and Mackey stress that it is important to bear in mind that “what learners say they do is not always the same as what they actually do” (Gass & Mackey, 2007, p. 45). The danger of participants reporting inaccurately is higher for retrospective than concurrent reports, as there is some time between the actual event and the verbal report. This is referred to as the veridicality problem (Russo et al., 1989). Such memory effects can be controlled for by conducting retrospective reports as closely as possible to the activity under scrutiny, and by providing participants with a stimulus (Bowles, 2010, p. 14; Sasaki, 2013, p. 6). Another potential threat to validity is the problem of reactivity – the danger that the method itself could alter cognitive processing (Russo et al., 1989). This is especially a concern for concurrent verbal reports, but is generally thought to be less of an issue in retrospective reports, as they are produced some time after the activity is finished and therefore influence cognitive processing and strategic behaviour to a lesser degree. However, there is still a reactivity problem of participants knowing that they are going to have to provide a report after the activity. Reactivity would also play a role if retrospective reports are conducted more than once throughout an activity, as in the studies on listening assessment outlined in the following. In addition to these potential threats to validity, researchers need to be aware of practical considerations when conducting verbal reports. Collecting and transcribing reports is very labour and time intense. Analysing verbal reports is not straightforward either, as the data needs to

be coded, ideally according to a coding scheme (Kasper, 1998) and by more than one researcher to calculate coder-reliability. These practical constraints affect sample size, so that usually only small populations can be studied, which makes it hard to generalise the findings.

Verbal reporting has been employed in a number of investigations on test taker cognition in listening assessment (Badger & Yan, 2012; Buck, 1991; Field, 2012, 2015; Harding, 2011; Holzknicht et al., 2017; Ockey, 2007; Wagner, 2008; Winke & Lim, 2014; Wu, 1998). In all of these studies retrospective verbal reports were used, and whenever the research involved longer listening passages they were broken up into shorter passages, followed by probe questions to initiate retrospection. As outlined above, breaking up the listening passages that way minimises veridicality problems, but at the same time might introduce issues of reactivity. In addition to probe questions, some of these researchers also used the written test items (Badger & Yan, 2012; Field, 2012, 2015; Harding, 2011) or the participants' eye-traces while they had been solving the items (Holzknicht et al., 2017; Winke & Lim, 2014) as stimuli to initiate retrospection. Such stimulated recalls (Gass & Mackey, 2000), as outlined above, further minimise problems of veridicality. All of the studies allowed students to self-analyse their thoughts, so the generated reports could be described as retrospective self-observations in Cohen's (1996) terminology. The authors of the studies report that the method yielded rich and insightful data and, in the case of Harding (2011) and Holzknicht et al. (2017), expanded upon the results of quantitative methods used.

3.4. Research design

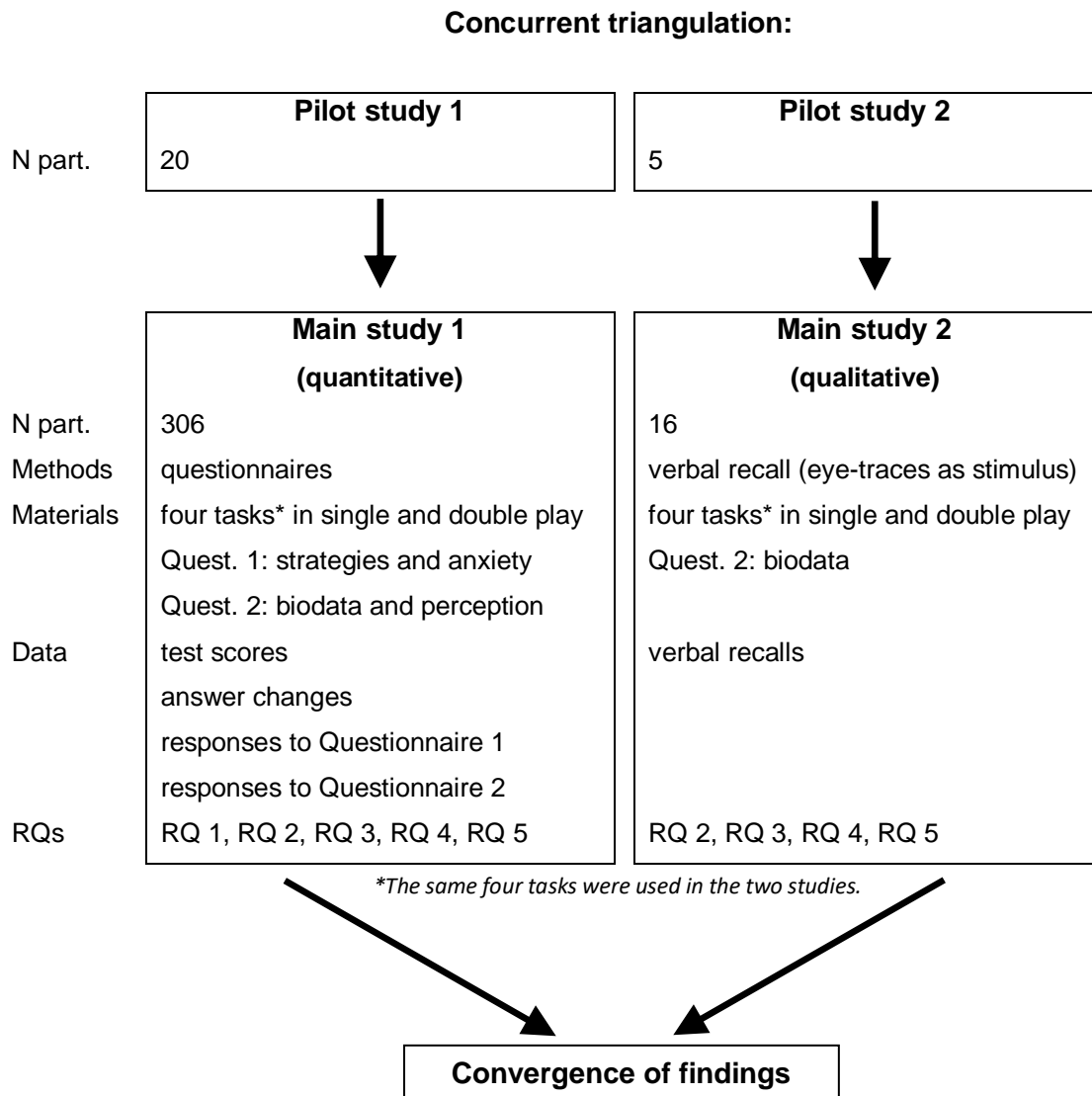
The research was framed in a mixed methods design employing the methods identified in Sections 3.2 and 3.3. Mixed methods designs have been used increasingly in language testing, in particular the combination of quantitative and qualitative methods, in order to be able to triangulate findings and thereby gain clearer insights into the complex nature of assessment (Jang et al., 2014). Thus, two studies were conducted in a concurrent triangulation design (Creswell, Plano Clark, Gutmann, & Hanson, 2003), utilizing the same listening tasks. In Study 1, 306 participants completed four listening tasks in a single play and double play condition and filled in two questionnaires. One questionnaire targeted listening strategies, test-taking strategies, and anxiety levels (Questionnaire 1) and the other questionnaire targeted biodata and different aspects

related to the tasks (Questionnaire 2). In Study 2, 16 participants completed the same tasks in both conditions and eye-tracking was used in combination with verbal recall to gain insights into test takers' cognitive processes, listening strategies, test-taking strategies and anxiety levels. Participants in Study 2 also filled in Questionnaire 2.

Two extensive pilot studies were conducted prior to data collection, one for each of the two main studies. 20 participants took part in the pilot for Study 1 and 5 participants piloted Study 2. The main aims of the pilot studies were a) to test the suitability of the different research methods to answer the research questions and b) to review the feasibility of the research design. For ease of reading, the findings of the pilot studies will not be presented separately first, but will instead be woven into the appropriate sections below.

The research design is summarised in Figure 5. The figure includes the number of participants for each part, the materials used, the type of data gathered and analysed, as well the research questions informed by the data. The detailed methodology of the two separate studies will be outlined below.

Figure 5: Overview of the research design



3.4.1. Ethical issues

Ethical consent was obtained from the Research Ethics Committee at Lancaster University prior to the two pilot studies. For both pilot studies, the participants' parents received an information sheet outlining the study and they were asked to sign a consent form for their children to be allowed to take part. For the main administration of both studies, the ethics documents were amended slightly so that students aged 16 or older were allowed to provide consent themselves without the need to consult their parents. This change was approved by the Research Ethics Committee at Lancaster University. In Study 1 all students received an information sheet in the lesson before the test and they signed a consent form on the day of the test. In Study 2 an information document

was sent to the students via email and they signed a consent form on the day of the experiment.

The information document and consent form addressed ethical issues relevant for the study. Participants were informed what the study entailed and what their role would be. They were also told that participation would be completely anonymous and that they would receive their test results some weeks after the administration. The document made it clear that the test results would not be used for classroom assessment and that the results were of no consequence to them. Participants were informed that they could choose not to take part in the study without any consequences for them, and to contact me should they decide not to participate. They were also told that they could reverse their decision to participate after they had completed the experiment, so that the data they provided would not be used. However, all of the participants I contacted chose to take part and none of them reversed their decision.

Participants were also informed that all of the data collected during the course of the research would be kept strictly confidential and that the results of the study would be used for academic purposes only. Any identifying information, such as names and other personal characteristics, are anonymised in the thesis or any other publications of this research. The data is also stored securely: All paper-based data is kept in a locked cupboard and electronic data is stored on an encrypted and password protected computer. The ethics documents, including the information sheet and consent form, are included in Appendix 1.

3.4.2. Research context

The research was carried out in the Austrian higher secondary school context. All participants were Austrian school students and the tasks used in the two studies were taken from past live papers of the English listening section of the standardised Austrian matriculation examination (Matura). Before the tasks are described in more detail, some context about the Matura as well as the development and structure of the listening test will be provided.

The Matura is a high-stakes test at the end of Austrian students' higher secondary education. It serves an important gatekeeping function in that students need to pass the Matura at the end of grade 8 in order to be able to study at University. For this reason, the Matura is developed by testing professionals at the Austrian Ministry of Education.

As described in detail in Spöttl, Eberharter, Holzknecht, Kremmel, and Zehentner (2018), the test development cycle for the foreign language exams in the Matura follows EALTA's guidelines for good practice in language testing (EALTA, 2006). It includes item writer training, item development based on standardised test specifications, item moderation, field trialling, statistical analyses based on field trialling, benchmarking and standard setting, as well as post-test analysis of the live administration.

Although the Matura is developed by language testing professionals, it is administered by the individual class teachers. During test administration, the class teachers need to adhere to detailed administration guidelines provided by the ministry. These guidelines include standardised instructions which need to be read out verbatim to students, as well as detailed information about seating arrangements or the preparation of audio equipment. The guidelines are described in more detail in Section 3.5.5 and in Appendix 4.

The individual class teachers also score the Matura exam. For the receptive skills, the ministry provides keys for closed-ended questions and detailed extended marking schemes for open-ended questions based on answers from the field trial. As students are not penalised for spelling mistakes in open-ended questions for the receptive skills, teachers need to decide whether an answer is correct or incorrect. To aid teachers in their decisions, they can consult a hotline and helpdesk service provided by the ministry on the days following the live test, whereby two language experts and two language testing professionals discuss the individual answers in plenary and come to a consensus decision for each answer (see Eberharter & Frötscher, 2012). Each acceptable and unacceptable answer is then entered into the extended marking scheme, which is made available to teachers online.

3.4.2.1. Matura listening tasks

The listening construct targeted in the Matura for academic upper secondary schools is based on the B2 listening descriptors of the Common European Framework of Reference for Languages³ (CEFR) (Council of Europe, 2001). All stages in test development, from item writer training to post-test analysis, are closely linked to the CEFR. In line with the B2 listening descriptors, the construct is also based on the

³ A B1 exam is also developed for a small number of students, however for reasons of relevance this will not be discussed here.

listening behaviours outlined in Green (2017, pp. 55–83). Accordingly, targeted behaviours in the Matura include listening for gist, listening for main ideas and supporting details, and listening for specific information and important details.

Each task generally consists of one coherent listening text based on authentic audio material. The conversations are not recorded by actors in a studio, but are either derived from authentic online sources or based on unscripted interviews with English speakers conducted by the item writers. The listening texts include monologues as well as dialogues and conversations between multiple speakers. A variety of standard English accents is targeted in the exam, including British English, American English, and Australian English. In terms of topics, item writers can choose from a range of personal, public, occupational, or educational domains as outlined in the CEFR, and they are encouraged to choose topics that are within the range of interests of 17-19 year old Austrian school students. Potentially distressing topics are avoided.

Three different test formats are developed for the B2 listening exam in the current version of the Matura: multiple-choice, note-form, and multiple-matching. Multiple-choice tasks consist of six to ten items, each of which includes a question or stem and four answer options with one correct answer. Note-form tasks are comprised of eight to ten items. For each of these items students need to either answer a question, fill in a gap in the middle of a sentence, or fill in a gap at the end of a sentence, but only one of these item types is used for each task. Students are not allowed to use more than four words for an answer. In multiple-matching tasks candidates need to match questions with answers or sentence beginnings with sentence endings. Each multiple-matching task consists of six to ten questions or sentence beginnings and corresponding matches, as well as two distractors. Regardless of the test format, all tasks include an example item at the beginning of the listening text.

The listening tasks for the foreign language exams are developed following the steps described by Green (2017), who trained the first item writers and test developers involved in task development. As outlined in detail in Spöttl et al. (2018), after choosing suitable listening texts according to the B2 listening descriptors in the CEFR, the trained item writers first map the texts in terms of essential information required for successful comprehension. This “textmapping” procedure (Sarig, 1989) varies slightly depending on the targeted listening behaviour (listening for gist, listening for main ideas and supporting details, or listening for specific information and important details), but is always based on a consensus of at least three out of four item writers. Next, the item

writers decide on a suitable task type for the text (multiple-choice, note-form, or multiple-matching) and develop items based on the “textmapping” results, followed by peer-feedback from other item writers and several loops of task moderation by professional language testing experts. The tasks are then piloted on a sample of at least 100 students from a population similar to the target population. Following rigorous statistical analyses of the field test data, including analyses of student questionnaires, tasks are either banked, revised and re-trialled, or eliminated. As a last step, successful tasks undergo a standard setting procedure, in which a panel of experts link each individual item of a task to the CEFR and set the cut score of the exam to ensure equal difficulty each year. The final exam for each language lasts about 45 minutes and includes four listening tasks of varying task formats, topics, accents, and targeted listening behaviours. All listening texts are played twice via a loudspeaker (headsets are not used). The exam is administered and scored by the class teachers following standardised marking schemes with all items being weighted equally. Table 2 shows a condensed set of specifications of the Matura B2 listening exam for English.

Table 2: Condensed set of specifications of the Matura B2 listening exam for English

Target level	CEFR B2
Test taker population	17-19 year old Austrian school students
Targeted behaviours	listening for: <ul style="list-style-type: none"> • gist • main ideas and supporting details • specific information and important details
Task formats	multiple-choice, multiple-matching, note form
Audio material	authentic (unscripted) monologues and conversations
Number of items per task	6-10
Number of tasks in live exam	4
Targeted accents	British English, American English, Australian English

3.4.2.2. Chosen tasks for the two studies

Four tasks from the English listening section of the Matura were chosen for the two studies. The tasks were standard set at CEFR level B2. Two of the tasks were multiple-choice (MC) tasks, where students had to choose one correct answer out of four. The other two were note-form (NF) tasks, where students had to fill in gaps at the end of sentences with a maximum of four words. Both of these task types are commonly used

in language assessment and also have the advantage of allowing the comparison of results with other research in this area, notably Field (2015). Initially, multiple-matching tasks were also piloted, however only a limited number of response processes could be identified in the stimulated recall data of the pilot study in relation to this task type. The pilot study yielded much richer results on MC and NF tasks in terms of verbal recalls, so it was decided to use MC and NF tasks for the main study. The following seven criteria were considered for selecting the tasks:

1. The task was used in a live-administration of the Matura, which guaranteed that it had passed all the quality control procedures described above.
2. The topics of the four tasks were different to avoid potential overlap in terms of topical knowledge.
3. The tasks targeted both standard British and American English in order to avoid overlap in terms of accent familiarity.
4. Tasks with the same format had the same number of items to allow for cross comparisons between item formats.
5. Both NF tasks were “fill in the blank at the end of sentences” format rather than “fill in the blank in the middle of sentences” or “answering questions” format to allow for cross comparisons (as described above, all three formats are developed for the Matura).
6. The tasks had similar task and item difficulty properties based on the field trial and standard setting results to allow for comparisons between tasks and item formats.
7. There was a variety of targeted listening behaviours to elicit both lower-order and higher-order cognitive processes from the test takers, however the targeted behaviour within each task type was the same to allow for cross comparisons across task types.

Table 3 summarises the tasks that were chosen for the two studies. One of the two MC tasks (“Apted’s film experiment”) was used in the live-administration of 2013. The task features an interview by a journalist from the US with the British TV director Michael Apted about a documentary he had just released. The other MC task (“Useful bottles”), developed for the live exam in 2012, is based on a recording featuring three British speakers about recycling plastic bottles into school uniforms. Both tasks included one example item and six items to answer, with four options and one correct

answer per item. The NF task “Swan upping”, about the tradition of looking after swans in British lakes and rivers, was also taken from the test booklet of the 2012 Matura exam. The second NF task (“Lego master model builder”) is a radio interview by a journalist from the US with a professional Lego builder from the US and was included in the 2015 Matura exam. Both NF tasks included one example item and nine items to answer.

Table 3: Summary of the tasks used in the two studies

Task ID	Task title	Format	Number of items	Audio file length	Mean FV trial	Std. setting	Target behav.*
MC1	Apted's film experiment	MC	6 + 1 example	4 min 01 sec	69%	B2	MISD
MC2	Useful plastic bottles	MC	6 + 1 example	3 min 41 sec	69%	B2	MISD
NF1	Swan upping	NF	9 + 1 example	3 min 40 sec	71%	B2	SIID
NF2	Lego master model builder	NF	9 + 1 example	2 min 50 sec	43%	B2	SIID

*MISD = main ideas and supporting details, SIID = specific information and important details

As can be seen in Table 3, the tasks were similar in terms of audio file length and mean item difficulty, as judged by facility values in the field trials, with the notable exception of NF2, which has a shorter audio file and is also somewhat more difficult. NF2 had to be included due to the limited number of tasks available which matched the criteria above. Despite these differences, the standard setting judges placed all four tasks at B2 level. Still, the higher difficulty of NF2 needs to be taken into account when interpreting the results.

In terms of targeted listening behaviour, the two MC tasks were developed to test main ideas and supporting details and the two NF tasks were aimed at testing specific information and important details.

3.5. Study 1

Study 1 was set up with the aim to explore differences between a single play and double play condition in terms of psychometric properties of listening test items (RQ 1), utilizing the listening tasks outlined in the last section. In addition, two questionnaires

on test takers' use of metacognitive strategies and anxiety levels (Questionnaire 1) as well as perceived task difficulty, perceived validity, and preferred task type (Questionnaire 2) were administered to all participants to inform RQ 2 on cognitive processes, RQ 3 on listening strategies, RQ 4 on test-taking strategies, and RQ 5 on anxiety. In this section, the study participants as well as the necessary task adaptations will be outlined first, followed by a detailed description of the two questionnaires. The section then goes on to presenting the research design and data collection, and it concludes with an outline of the analysis procedure.

3.5.1. Participants

As outlined in Section 3.4.2, the target population of the B2 tasks used in the study are typically 17-19 year old students in grade 8 of an Austrian academic upper secondary school. However, according to the Austrian academic upper secondary curriculum, students should already have reached B2 at the start of grade 7. It was therefore decided to recruit students from grade 7 instead of grade 8 to take part in the study, for two main reasons:

1. Test takers in grade 8 might have already been familiar with the tasks, as in grade 8 teachers often use tasks from past Matura papers in class in order to prepare students for the school leaving exam (all of the tasks used in the two studies were in the public domain at the time of data collection).
2. It was hoped that students in grade 7 would find the tasks slightly more challenging than students in grade 8, which would potentially yield richer and more insightful data as students would need to display a greater amount of controlled processing and therefore be able to recall their thoughts more accurately (see Green, 2017, pp. 3–5).

A typical grade 7 secondary school student in Austria will have first encountered English in primary school at age 6, and by grade 7 in upper secondary school will be studying for up to five hours per week in class. After primary school, students typically start learning a second foreign language in grade 1 of secondary school at age 10, however this varies from school to school and also from class to class within schools.

Students were recruited via their class teachers, who were known to me through professional networks. In total, 306 students (197 female and 109 male) attending 16

different grade 7 classes took part in the study. The classes were spread across five academic upper secondary schools over three regions across Austria (Upper Austria, Styria, and Vorarlberg).

The majority of students were 16 or 17 years old (69.9% and 26.5% respectively, see Table 4), with a smaller number of students aged 18 (3.6%). German was the L1 of most of the participants (91.2%, as shown in Table 5). Thirteen percent of the students grew up bilingually, with 13 students (4.2%) having English as a second L1 (Table 6).

Table 4: Study 1: participants' age

Age	N	%
16	214	69.9
17	81	26.5
18	11	3.6
total	306	100.0

Table 5: Study 1: participants' L1

L1	N	%
German	279	91.2
French	1	0.3
Italian	1	0.3
Turkish	4	1.3
Serbian	5	1.6
other	8	2.6
missing	8	2.6
total	306	100.0

Table 6: Study 1: participants with a second L1

Second L1	N	%
English	13	4,2
French	4	1,3
Spanish	2	0,7
Italian	3	1,0
Turkish	6	2,0
Croatian	2	0,7
Serbian	2	0,7
Hungarian	2	0,7
other	6	2,0
total	40	13,1

3.5.2. Task adaptations

For Study 1 the original pen-and-paper Matura versions of the four tasks were used, with slight changes in the instructions (Figure 6). One change related to the two conditions (single play and double play): The information about how many times the recording would be played was added to the instructions. The number of times the recording would be played was also underlined in the written instructions, as it was found in the pilot study that some test takers had missed this information. It was important that candidates knew from the outset how many times they would hear the recording, in order to be able to compare the single play with the double play condition. In any real-life listening test candidates would also be provided with this information, and it was hypothesised that they would approach and complete a task differently if they knew from the start that they would hear the recording once compared to twice.

Another change related to the amount of time given to test takers to complete the tasks in the single play condition. In the Matura exam (which utilises double play, as outlined in Section 3.4.2.1) participants are given 45 seconds to study the task before the listening text starts, 15 seconds after the first play to check their answers before the second play starts, and another 45 seconds after the listening text is finished to finalise their answers. Study participants were given the same amount of time as in the Matura tasks in the double play condition. However, in the single play condition, participants were given 60 seconds after the listening text finished to finalise their answers, as it was found in the pilot study that 45 seconds were not sufficient for students to be able to note down all responses in the NF tasks. Thus, the time available to preview and answer questions was eventually the same in the two conditions: 45 seconds preview for both conditions, 15 seconds + 45 seconds to answer questions in double play, and 60 seconds to answer questions in single play.

As in the original Matura tasks, the instructions were also simultaneously delivered on audio by a female speaker. Due to the different question numbering (in the original test booklet the tasks appeared in different positions so the question numbers were different), the different number of times the listening text was played, and the difference in time given to complete the tasks in the single play condition, the instructions for each task in each condition had to be re-recorded for the study. The re-editing of the recordings was done using the software Audacity (version 2.1.2 for Mac),

which allowed me to insert the new instructions as well as the appropriate pauses between instructions and listening text in a standardised way.

Figure 6: Example of task instructions

You are going to listen to a recording about a documentary TV series made by Michael Apter. First you will have 45 seconds to study the task below, then you will hear the recording once. While listening, choose the correct answer (A, B, C or D) for questions 1-6. Put a cross (☒) in the correct box. The first one (0) has been done for you.

After listening, you will have 60 seconds to check your answers.

3.5.3. Questionnaire 1: strategies and anxiety

In addition to completing the four tasks, students also filled in a questionnaire targeting listening strategies and test-taking strategies to inform RQ 3 and RQ 4 as well as test-taking anxiety and listening anxiety to inform RQ 5 (see Appendix 2). The questionnaire was constructed following the guidelines by Dörnyei and Taguchi (2009, pp. 127–128) and consisted of 25 statements to which participants had to indicate their level of agreement on a four-point Likert scale (disagree, partly disagree, partly agree, agree). They could also choose “I don’t know”. The questionnaire was administered in German but is translated in Appendix 2 for ease of reading.

The statements on test-taking anxiety (7 to 13) were taken from Winke and Lim (2014), who based their questionnaire on Cassady and Johnson (2002). The statements on listening anxiety (20-25) were adapted from Elkhafaifi (2005), also used by Brunfaut and Révész (2015). The items on test-taking strategies (1-6) were based on Cohen and Upton (2007) and adapted by Winke and Lim (2014), and the items on listening strategies (14-19) were based on Vandergrift (1997) and adapted by Winke and Lim (2014).

A number of considerations guided the development of the questionnaire used in this study. The original questionnaires contained more statements than could be administered in the study, so only those statements which were considered most relevant for the purpose of the study were included. The decision on whether a statement should be included in the study was taken by me based on three factors. First, a number of statements were phrased in very general terms, particularly in Elkhafaifi’s questionnaire on listening anxiety (e.g. “I am worried about all the new sounds you have to learn to understand spoken [English]”), but also in the other questionnaires (e.g. “I am less

nervous about tests than the average college students” in Winke and Lim’s questionnaire on test-taking anxiety). Such general statements were not included in the study, as it would have been difficult to relate these statements to the specific tasks the students had just performed in a single play or double play condition. Second, a number of statements would not have been relevant in relation to the tasks the students had just performed (for example, the statement “You have to know so much about [English] history and culture in order to understand spoken [English]” in Elkhafaifi’s questionnaire on listening anxiety). Finally, some statements overlapped with other statements and were therefore omitted. For example, the item “While listening, I ignore irrelevant information” in the original questionnaire on listening strategies overlapped with “I only listen for relevant information to answer the questions” in the original questionnaire on test-taking strategies, so only the latter was included.

Participants replied to the questionnaire twice: once after completing two tasks (of the same format) in a single play condition and again after completing the other two tasks (of the other format) in a double play condition (the rationale behind this is described in the research design in Section 3.5.5 below). It was found in the pilot study that it was not always clear for participants what the statements refer to: whether they refer to their listening test experience in general or to the tasks they had just completed specifically. For this reason, in the main study all statements were phrased in past tense and detailed instructions were included to make it clear to the participants that they should relate their answers only to the two tasks they had just completed.

Two statements varied slightly between the two task formats used: “I read the answer options before listening” (Item 1) in relation to the MC tasks was changed to “I read the questions before listening” in relation to the NF tasks and “I listened for the words that appeared in the questions and options” (Item 4) in relation to the MC tasks was changed to “I listened for the words that appeared in the questions” in relation to the NF tasks. Also, Item 2 (“I predicted my own answer after listening and then looked at the options”) was not included in the NF version as it was not relevant for this task type.

3.5.4. Questionnaire 2: biodata and task perception

After completing the experiment participants filled in another questionnaire including items on gender, age, L1, and, following recommendations by Alderson, Clapham, and

Wall (1995), questions about the tasks (see Appendix 3). It is important to take test takers' perspectives into account in validation research, as they may not perform to the best of their abilities if the test appears to be measuring skills other than the ones stipulated in the construct, or if the test appears too difficult (Schmitt, 2002; Xie, 2011). Such factors could introduce construct-irrelevant variance into test scores. Therefore, the questionnaire asked participants about their familiarity with the topics and the task types, their perceived difficulty of the tasks, and how well they were able to show their listening competency in each of the four tasks. The final question asked whether participants preferred single play or double play and for what reasons. The questionnaire was administered in German but is translated in Appendix 3.

3.5.5. Research design and procedure

The 306 participants completed the tasks and filled in the questionnaires in a complex and carefully counter-balanced design. I developed 16 different versions of the test and counter-balanced them across groups of participants to control for potential confounding factors. As shown in Table 7, in version 1 test takers first completed both MC tasks in a double play condition, followed by the questionnaire targeting strategic behaviour and anxiety (Questionnaire 1). They then completed both NF tasks in a single play condition, followed again by Questionnaire 1. The questionnaire was administered twice, as it was hypothesised that test takers would respond differently depending on the condition they had just experienced (single play or double play). Therefore, the same task type was used within each condition across all versions, in order not to confound questionnaire responses with potential task type effects. For example, test takers might react differently to a single play condition for MC tasks than to a single play condition for NF tasks in terms of test-taking strategies, listening strategies, test-taking anxiety, or listening anxiety. If the two different task types had been used within the same condition, such differences would not have been captured by the questionnaire responses, and may have weakened the validity of the responses themselves. After completing Questionnaire 1 for the second time, test takers filled in the biodata and task perception questionnaire (Questionnaire 2).

Table 7: Study 1: research design

# of times heard	Version 1	Version 2	Version 3	Version 4
2	MC 1	MC 1	MC 2	MC 2
2	MC 2	MC 2	MC 1	MC 1
	Questionnaire 1	Questionnaire 1	Questionnaire 1	Questionnaire 1
1	NF 1	NF 2	NF 1	NF 2
1	NF 2	NF 1	NF 2	NF 1
	Questionnaire 1	Questionnaire 1	Questionnaire 1	Questionnaire 1
	Questionnaire 2	Questionnaire 2	Questionnaire 2	Questionnaire 2
	Version 5	Version 6	Version 7	Version 8
2	NF 1	NF 1	NF 2	NF 2
2	NF 2	NF 2	NF 1	NF 1
	Questionnaire 1	Questionnaire 1	Questionnaire 1	Questionnaire 1
1	MC 1	MC 2	MC 1	MC 2
1	MC 2	MC 1	MC 2	MC 1
	Questionnaire 1	Questionnaire 1	Questionnaire 1	Questionnaire 1
	Questionnaire 2	Questionnaire 2	Questionnaire 2	Questionnaire 2
	Version 9	Version 10	Version 11	Version 12
1	MC 1	MC 1	MC 2	MC 2
1	MC 2	MC 2	MC 1	MC 1
	Questionnaire 1	Questionnaire 1	Questionnaire 1	Questionnaire 1
2	NF 1	NF 2	NF 1	NF 2
2	NF 2	NF 1	NF 2	NF 1
	Questionnaire 1	Questionnaire 1	Questionnaire 1	Questionnaire 1
	Questionnaire 2	Questionnaire 2	Questionnaire 2	Questionnaire 2
	Version 13	Version 14	Version 15	Version 16
1	NF 1	NF 1	NF 2	NF 2
1	NF 2	NF 2	NF 1	NF 1
	Questionnaire 1	Questionnaire 1	Questionnaire 1	Questionnaire 1
2	MC 1	MC 2	MC 1	MC 2
2	MC 2	MC 1	MC 2	MC 1
	Questionnaire 1	Questionnaire 1	Questionnaire 1	Questionnaire 1
	Questionnaire 2	Questionnaire 2	Questionnaire 2	Questionnaire 2

As shown in Table 7, the test was administered in 16 different versions (one for each class) to control for potential ordering effects, so 16 different test booklets with corresponding audio files and questionnaires were developed. Questionnaire 2 was also developed in 16 different versions, as all four tasks were listed for three questions and the order of the tasks changed for each version (see Appendix 3).

Then, each participating class was assigned to take one of the 16 versions of the test, so all individuals within a class took the same version of the test. Due to this research design, the participants were divided into two sub-groups. All participants from sub-group 1 took the MC tasks in a double play condition and the NF tasks in a single play condition (versions 1-4 and 13-16) and participants from sub-group 2 took the MC tasks in a single play condition and the NF tasks in a double play condition (versions 5-12).

All materials were sent by post to the students' teachers and they were given a month to conduct the test and send the materials back to me. During this month the test and questionnaires were administered in a 50-minute session by the teachers. The

teachers received detailed instructions for test administration and were asked to strictly adhere to these instructions. The instructions were based on the (unpublished) test administration documents of the Austrian Matura exam (see Section 3.4.2) and on my experiences during the pilot study. They included information on how to prepare the room prior to the administration, describing in detail how tables should be arranged to minimise the chance for cheating. Each participant was allocated a number which corresponded to the number on a stick-on label, which was stuck at each seating place. The document also contained instructions on how the audio equipment should be prepared. The instructions for the participants had to be read out verbatim to maximise standardization across administrations. The complete instructions document (in German) is included in Appendix 4.

On the day of the test, participants were told in detail what they had to do and that they should treat the test like a normal classroom assessment. During the test, the teachers filled in a seating plan (Appendix 5), which allowed me to identify whether candidates had enough space to prevent cheating. The teachers also filled in a test administration report (Appendix 6) in which they were asked to specify whether there were any problems during the test, such as corrupted audio files, loud noises which might have impacted the candidates' understanding of the audio files, mistakes in the test booklets, candidates who came too late or had to leave earlier, candidates who behaved inappropriately, or specific questions from candidates.

In addition to the test booklets, candidates were also given two pens in a different colour (blue and red) and they were instructed by the class teacher to use the red pen only during the second play in the tasks which were played twice. The instructions in the audio file also included this information: Before each second play (after the 15 seconds between the two plays, see Section 0) the female speaker said "Use the red pen now" and before the following task (after the 45 seconds students were given to check their answers) she said "Use the blue pen again". This was done to be able to analyse how participants changed their answers during the second play. It was hoped that this analysis would inform RQ 3 on listening strategies and RQ 4 on test-taking strategies. Participants could keep the pens and also received their individual results some weeks after the administration.

3.5.6. Analysis

Prior to data analysis I checked the test administration reports and seating plans to identify candidates who might need to be excluded. Apart from one candidate using a white-out pen after the first play of the two double play tasks, all administrations ran as expected without any noteworthy problems. I also checked whether participants who had English as a Second Language performed differentially, which was not the case, so all candidates were included in the data analysis.

Next, I scored the 306 test booklets. All items in the single play condition were scored dichotomously as either correct or incorrect. For the MC tasks the published keys were used. For the NF tasks the extended marking schemes were obtained from the Austrian Ministry of Education. As described in Section 3.4.2, in the Matura exam students are not penalised for spelling mistakes in their answers to NF questions. Therefore, the exam developers have set up a hotline and helpdesk service during the live exam in which teachers can inquire about whether an answer is acceptable or not. Two language experts and two language testing professionals discuss the individual answers in plenary and come to a consensus decision for each answer (see Eberharter & Frötscher, 2012). Each acceptable and unacceptable answer is then entered into the extended marking scheme. The items completed in the double play condition were scored twice according to the keys and the extended marking schemes: once for answers after the first listening (as indicated by the blue pen) and once for answers after the second listening (as indicated by the red pen). The second scoring included a total of nine scoring categories, which are described in more detail in Section 4.3. Only one candidate had to be excluded from this analysis, as they used a white-out pen after the first play (see discussion above). For this candidate, only their final answers were scored.

After scoring I analysed the data in three stages. First, utilizing CTT (see Section 3.2), item statistics and reliability indexes were calculated in SPSS (version 24 for Mac) for each task and condition to identify how they were impacted by single play as compared to double play. Second, utilizing IRT, a bias analysis was performed using Many-Facet Rasch Measurement (MFRM, Linacre, 1994) to detect potential differences in task difficulty across the two conditions and task types. Finally, the extent to which students changed their answers during the second play as indicated by the use of the blue and red pens were analysed in detail.

The data for the two questionnaires were analysed in two steps. For Questionnaire 1, I calculated the mean and standard deviation for each statement first and, following that, performed a factor analysis and subsequent Wilcoxon signed-rank test to explore potential statistical differences in test takers' strategic behaviour and anxiety levels between the two conditions. For Questionnaire 2, frequencies were calculated and the responses to the open question were categorised and coded using the qualitative data analysis software Atlas.ti (version 8.4 for Mac).

For ease of reading, the analysis procedures for each stage will be outlined in more detail in the relevant results sections.

3.6. Study 2

In Study 2 retrospective and stimulated recall was used in combination with eye-tracking to track participants' response processes while they were completing the tasks in a single play and double play condition to inform RQ 2 on cognitive processes. It was hoped that the findings would also be informative with regards to test takers' strategic behaviour (RQ 3 and RQ 4) and anxiety levels (RQ 5). The tasks used were the same as in Study 1 to allow for cross-comparisons and triangulation.

Initially, a free recall procedure was also piloted with two participants. However, a preliminary analysis of the pilot study data based on idea units revealed that due to the length of the listening texts used in the study, memory effects impacted the completeness of the recalls substantially. I felt that collecting several free recalls at shorter intervals during task completion to counter such memory effects would have interrupted the test-taking process considerably and would have led to reactivity problems. A second pilot using retrospective recall and stimulated recall based on eye-movements yielded much richer results and was also less restrictive in terms of potential reactivity effects. I therefore decided not to use free recall in the main study, but to use retrospective and stimulated recall instead.

In the following, the study participants and adaptations to the tasks will be presented first, followed by a detailed outline of the research design and data collection procedure. In addition, the three steps during the analysis are described in detail, including transcription of the verbal recalls, coding of the data, and the double-coding process.

3.6.1. Participants

As in Study 1, students attending a 7th grade in an Austrian academic upper secondary school were recruited. The students were contacted through English teachers at their school. The teachers informed their class about the opportunity to come to the University of Innsbruck to take part in a research project involving English listening tests and eye-tracking. Due to the proximity to the University I only approached teachers from schools in Innsbruck.

In total, 23 students were interested in the study and I contacted all of them by email. Eighteen students replied and came to the University to perform the experiment; however, two students were not able to take part in the experiment on the day as they were wearing thick eyeglasses which led to insufficient eye-tracking readings.

The final sample consisted of 16 participants aged 16 (9 participants) or 17 (7 participants) attending grade 7 in three different academic upper secondary schools in Innsbruck. Nine participants were female and seven were male. All participants were German native speakers, with two participants having had extended exposure to English at some period in their life (one participant attended a bilingual kindergarten and primary school and another participant had lived in England for two years as a child). Two participants had a second L1, speaking German and Italian and German and Polish, respectively.

3.6.2. Task adaptations

As in Study 1, the original Matura task instructions had to be changed in relation to item numbering, condition (single play or double play), and extra time given for the single play condition (see Section 0), however, for Study 2 the tasks themselves also had to be adapted from the original pen-and-paper to html format for use on a computer and eye-tracker. While the general layout of the tasks including font and colouring remained the same, three aspects had to be changed. First, the instructions for each task were shown on a separate page which appeared on screen before the actual tasks. Second, for the MC tasks the layout and arrangement of items was changed slightly for the eye-tracking experiment to be able to fit all items on the screen without the need for scrolling (Figure 7). As in the original pen-and-paper Matura test, test takers were able choose more than one answer for each question (although only one answer was correct) or could leave the answers blank. Participants used the mouse to choose an answer.

Figure 7: Study 2: layout of MC tasks

Apted's film experiment

Q0 Michael Apted's film experiment looks at

A. the life of 7-year-old children in different times. ☐

B. how friendships change over time. ☐

C. 7 years in various children's lives. ☐

D. the development of some children over time. ☒

Q3 If this film experiment had been produced later, it would have turned out

A. quite different altogether. ☐

B. exactly the same. ☐

C. different in some details. ☐

D. only different in style. ☐

Q6 From age 7 to 49, Neil's life was

A. marked by mental illness. ☐

B. on a steady downward slide. ☐

C. a life of ups and downs. ☐

D. a life full of misery. ☐

Q1 The aim was to find out if people's lives in the UK are determined by

A. social background. ☐

B. gender. ☐

C. education. ☐

D. personality. ☐

Q4 Later in life, one horrible 7-year-old

A. became a famous politician. ☐

B. had serious mental problems. ☐

C. separated from her family. ☐

D. changed for the better. ☐

Q2 Apted's film "49 Up"

A. has only just come out. ☐

B. is nominated for an award. ☐

C. is a remake of an old film. ☐

D. starts a new reality TV show. ☐

Q5 This film experiment makes Apted wonder

A. why people feel so miserable about life. ☐

B. if all people are, in fact, equally interesting. ☐

C. whether making those films was right. ☐

D. how we can help people to cope with life. ☐

Third, although the arrangement of the NF tasks was not changed for the eye-tracking experiment, the tasks were programmed so that most of the eye-tracking screen was used (Figure 8). Participants could use the mouse or the keyboard to navigate between answer spaces and they had to use the keyboard to type their answers. As in the original pen-and-paper version, participants could leave answer spaces blank and they were not restricted in their number of words to answer the questions.

Figure 8: Study 2: layout of NF tasks

Swan upping

0	The main task of the swan uppers is...	conservation
1	In the past, swans were...	
2	A big trouble for swans is...	
3	Swan uppers talk to school children in order to avoid...	
4	Six teams of swan uppers in six boats wear...	
5	Prof. Perrins makes sure that the number of swans is...	
6	During swan upping, the swans are surrounded by the boats, then... <small>(Give <u>one</u> answer.)</small>	
7	On the riverbank, the swans are... <small>(Give <u>one</u> answer.)</small>	
8	When the swan uppers pass Windsor castle, they ... the Queen.	
9	All swans in Great Britain belong to...	

In addition to these changes, the audio file of each task was linked to the html file, so that the audio file start times were standardised across all participants. The audio file started as soon as the task appeared on screen, with 45 seconds of silence for participants to study the task (see Section 0). This was important because different audio file start times would have impacted eye-movement readings, as participants would have looked at the screen for a different amount of time.

3.6.3. Research design and procedure

The 16 participants each completed two tasks of the same task type on a Tobii TX300 eye-tracker. In terms of fixation filter, the default settings of the Tobii I-VT filter were used (see Holmquist et al., 2011). The tasks were the same as in Study 1. One task was administered in a single play condition and the other task in a double play condition. The same task type was used across the two conditions to be able to compare test taker processes between conditions. To balance the design and control for potential ordering effects, eight different test versions were programmed so that a maximum of two participants completed each version (Table 8).

The participants came to the University individually in order to perform the experiment. Before the test I told the participants in detail what they had to do. They were asked to treat the experiment like a normal classroom assessment. They then performed an eye-tracking calibration task to find the optimal seating position for accurate eye-tracking readings, followed by an example task to get used to the screen layout and set-up. The example task consisted of the first two questions (and one example) from another Matura listening task in the same task format. After the example task any unanswered questions were addressed. Another calibration was run before the experiment was started with the first task. Participants completed the first task without interruption, either in a single play or double play condition, depending on which version they took (see Table 8). They knew from the outset whether they were going to hear the recording of a task once or twice, as this was stated in the instructions.

Table 8: Study 2: research design

# of times heard	Version 1	Version 2	Version 3	Version 4
	example MC task	example NF task	example MC task	example NF task
1	MC 1	NF 1	MC 2	NF 2
	retrospective recall	retrospective recall	retrospective recall	retrospective recall
	stimulated recall	stimulated recall	stimulated recall	stimulated recall
2	MC 2	NF 2	MC 1	NF 1
	retrospective recall	retrospective recall	retrospective recall	retrospective recall
	stimulated recall	stimulated recall	stimulated recall	stimulated recall
	post-hoc interview	post-hoc interview	post-hoc interview	post-hoc interview
	Questionnaire 2	Questionnaire 2	Questionnaire 2	Questionnaire 2
# of times heard	Version 5	Version 6	Version 7	Version 8
	example MC task	example NF task	example MC task	example NF task
2	MC 1	NF 1	MC 2	NF 2
	retrospective recall	retrospective recall	retrospective recall	retrospective recall
	stimulated recall	stimulated recall	stimulated recall	stimulated recall
1	MC 2	NF 2	MC 1	NF 1
	retrospective recall	retrospective recall	retrospective recall	retrospective recall
	stimulated recall	stimulated recall	stimulated recall	stimulated recall
	post-hoc interview	post-hoc interview	post-hoc interview	post-hoc interview
	Questionnaire 2	Questionnaire 2	Questionnaire 2	Questionnaire 2

After the first task was completed (i.e. after one play in the single play condition or two plays in the double play condition), participants were asked to recall their thoughts during task completion. Based on the literature review, it was decided to collect both retrospective recalls without a stimulus (Buck, 1991; Ockey, 2007; Wagner, 2008; Wu, 1998) as well as stimulated recalls (Badger & Yan, 2012; Brunfaut & McCray, 2015; Field, 2012, 2015; Harding, 2011; Holzknecht et al., 2017; McCray et al., 2012; Winke & Lim, 2014). It was hoped that retrospective recalls without a stimulus would encourage test takers to reflect on their thoughts in more general terms, while stimulated recalls would re-activate more specific thought processes that occurred during task completion. Thus, the participants were first asked general questions about their thoughts during the test (retrospective recalls) following the standardised prompts outlined in Section 3.6.4 below. They then watched a recording of their eye-movements while they had been solving the items, overlaid with the audio of the task, to stimulate further recalls (Holzknecht et al., 2017; Winke & Lim, 2014). Test takers could stop the recording at any time to recall their thoughts. In addition, I stopped the recording whenever I noticed unexpected eye-movements in the recording (e.g. when the participant had focussed on one word or a particular answer for a long time) or when the test taker showed obvious reactions such as laughing or nodding. The recording was also stopped when participants remained silent for long stretches, however this was not necessary in the great majority of cases. After finishing the recalls on the first task, test

takers performed another eye-tracking calibration before they completed the second task without interruption. Following task completion, another session of retrospective and stimulated recalls was conducted. Finally, after completing the recalls for the second task, participants were asked three general questions about the experiment (see Section 3.6.4 below). As recommended by Bowles (2010), participants were encouraged to use German or English (or a combination of the two) in their recalls (see Bowles, 2010, p. 115). The recalls were recorded via a high-quality table microphone, as well as a video recording of the entire procedure which served as backup. Each session, including the instructions and example task, took about 1.5 to 2 hours, depending on how extensively participants were able to recall their thoughts. Participants could take a break after completing the stimulated recall of first task.

After the experiment the participants filled in Questionnaire 2 (see Section 3.5.4). For the eye-tracking experiment, Questionnaire 2 only included questions about the two tasks the participants had just performed. The 16 students who participated in the experiment were compensated with 10€ for their time. The two students who could not perform the experiment due to their eyeglasses were compensated with 5€ for taking the time to come to the University.

3.6.4. Prompts for verbal recall

The following standardised prompts were used for the verbal recall sessions after each task. The original prompts were informed by previous studies on test takers' response processes in listening assessment (Harding, 2011; Holzknecht et al., 2017) and were refined and slightly adapted based on findings from the pilot study. The prompts were in German but are translated here for ease of reading.

Retrospective recall

I am now going to ask you some questions.

What were you thinking while you were listening to the text and working on the task?

Did you experience any difficulties while listening to the text and working on the task?

YES → *Why was it difficult?*

NO → *Why was it not difficult?*

How were you listening?

Stimulated recall

You are now going to see a video of your eye-movements while you were working on the task. Look at the video and try to remember what you were thinking. I am interested in what was going on in your head while you were listening and working on the task. You can stop the video at any time by clicking the “pause” button here. Stop the video as soon as you remember what you were thinking. You can stop the video as often as you wish. You can’t make any mistakes and anything you say is useful.

Pre-listening (i.e. before the listening text started)

Participant pauses and recalls thoughts. Look at eye-movements and ask specific questions at the end, e.g.

You did (not) read all of the questions/answers. Why (not)?

You focused on this question for a long time. Why?

While-listening (i.e. while the listening text was played)

If the participant does not stop the recording to recall their thoughts regularly enough, pause the recording at the specified points and ask participants the following questions:

- After about 10 seconds:
 - *What were you thinking while you were listening to the recording?*
 - *How were you listening?*
- Particular reactions (e.g. unexpected eye-movements, laughter, nodding, etc.)
 - *Here you focus on X / You are laughing/nodding etc. Can you tell me more?*
What were you thinking while you were listening?
 - *How were you listening?*
- At the end of the task
 - *What were you thinking when you listened to the end of the recording?*
 - *How were you listening?*

Post-listening (i.e. after the listening text finished playing)

What were you thinking when the recording was finished?

Post-hoc interview

How did you find the experiment?

What do you prefer, single play or double play? Why?

Was the video of your eye-movements useful or distracting for helping you remember your thoughts? Why?

3.6.5. Analysis

Prior to data analysis I checked whether the two participants who had had extended exposure to English at some period in their life (see Section 3.6.1) outperformed the other students, which may have skewed the results. This was not the case, so all 16 participants were included in the analyses.

The verbal data was then analysed in three steps. First, all of the recalls were transcribed by me with the help of a research assistant. Then, I coded the recalls according to the coding scheme, after which two additional coders applied the same coding scheme to 20 percent of the data to check the reliability of the coding process. Each of these steps will be outlined in more detail in the following.

Initially I also planned to analyse eye-tracking metrics, however based on experiences during data collection this avenue was not explored. During the eye-tracking experiments I noticed that participants would sometimes look at a certain word for an unusually long time. When I asked them about this during the stimulated recalls, participants often stated that they did not actually pay attention to what they were looking at, but were only focussing on understanding the listening text. Thus, it would have been impossible to disentangle candidate's reading processes from their listening processes through eye-tracking data, so I only focussed on the verbal report data.

3.6.5.1. Transcription

I transcribed the first two verbal recalls using the audio recordings (rather than the video recordings) and by so doing I developed detailed transcription conventions. Then, a research assistant transcribed the remaining 14 recalls from the audio recordings following the transcription conventions, under my close supervision. The research assistant used the transcription software F4 (version 6.2.6 for Windows), which links the transcription with the audio file and inserts timestamps at each line break, so I was able to quickly locate passages in the audio files during the regular checks. The data was transcribed in the language used by the participants (i.e. German with occasional English phrases) to avoid data loss due to translation into English. Only the data used for illustrative purposes in the thesis were translated by me.

I checked transcription accuracy and adherence to the transcription conventions as soon as the research assistant finished the transcription of a participant's recall. I read through the transcription in detail and added missing information by double-checking the audio recording. If I did not understand a missing passage from the audio recording either, I checked the video recording for further clarification.

Once a transcription was checked, I segmented each verbal recall into 10 different data files according to task type, task, number of times a candidate listened to the recording, and stage of recall. This allowed for more efficient comparison of the verbal recall data across the different task types, tasks, and conditions. Table 9 presents an illustration of the 10 data files, based on the verbal recall data from Participant 1 (who completed MC1 in single play and MC2 in double play). The first file contains all data from the retrospective recall performed after the participant completed MC1 in single play (see also Sections 3.6.3 and 3.6.4). The next three data files contain the stimulated recall data (with participant's eye-traces overlaid with the audio file functioning as a stimulus). The first of these (P01 MC1 once pre-listening), contains all data drawn from the pre-listening period: the 45 seconds during which the participant could study the task before listening. The next file (P01 MC1 once task) includes the stimulated recall data drawn from the while-listening period: the time during which the participant listened to the text and engaged with the task. The third file contains data from the 60 seconds in which the participant could finalise their answers (P01 MC1 once post-listening). The transcription of the second task (MC2) was split up in the same way, but because this task was completed in the double play condition, the stimulated recalls were also labelled according to the number of listening (first or second) and they were segmented accordingly into two additional files: one file containing all data from the while-listening period during the second play (P01 MC2 twice task second) and another file including the data from the post-listening period (the 45 seconds the students could finalise their answers after the second play - P01 MC2 twice post-listening second). Finally, the post-hoc interview was also saved as a separate data file (P01 post-hoc). Appendix 8 includes exemplary excerpts from the retrospective recalls, the stimulated recalls, and the post-hoc interview.

Table 9: Study 2: separate data files and types of data for Participant 1

Document name	Retrospective recall	Stimulated recall	Post-hoc interview
P01 MC1 once retrospective	X		
P01 MC1 once pre-listening		X	
P01 MC1 once while-listening		X	
P01 MC1 once post-listening		X	
P01 MC2 twice retrospective	X		
P01 MC2 twice pre-listening first		X	
P01 MC2 twice while-listening first		X	
P01 MC2 twice while-listening second		X	
P01 MC2 twice post-listening second		X	
P01 post-hoc			X

3.6.5.2. Coding

Once all transcripts were checked by me and segmented into different data files as outlined in Table 9, the coding scheme was finalised. The scheme was based on the theoretical framework described in Section 2.4 and included the four main response processes of interest: cognitive processes, listening strategies, test-taking strategies, and anxiety. The main aim of the research in this thesis was to identify how these dimensions are influenced by double play as compared to single play. In the following, a summary of the four response processes and their various sub-dimensions will be presented, based on the discussion in Section 2.4. This summary served as the first version of the coding scheme.

Cognitive processes

Descriptions are based on Field (2013), Vandergrift and Goh (2012), and Rost (2011)

- Input decoding (acoustic-phonetic processing): Recognising incoming sounds as speech. Informed by phonological knowledge.
- Lexical search: Recognizing individual words. Informed by lexical knowledge.
- Parsing: Putting individual words into a syntactic pattern to form the bare meaning of an utterance at clause or sentence level. Informed by syntactic knowledge.
- Meaning construction (micro-level conceptualization): Relating the literal meaning of utterances to the context in which they occurred to construct higher-level meaning. Informed by pragmatic knowledge and external knowledge about the world, the speaker, and the topic.

- Discourse construction (macro-level conceptualization): Relating the meaning of the message to the discourse as a whole. Informed by external knowledge about the text type, the world, and the speaker.

Listening strategies

Descriptions are taken verbatim from Vandergrift and Goh (2012, pp. 277–284)

- Planning: Developing awareness of what needs to be done to accomplish a listening task, developing an appropriate action plan and/or appropriate contingency plans to overcome difficulties that may interfere with successful completion of a task.
- Focusing attention: Avoiding distractions and heeding the auditory input in different ways, or keeping to a plan for listening development.
- Monitoring: Checking, verifying, or correcting one's comprehension or performance in the course of a task.
- Evaluation: Checking the outcomes of listening comprehension or a listening plan against an internal or an external measure of completeness, reasonableness, and accuracy.
- Inferencing: Using information within the text or conversational context to guess the meanings of unfamiliar language items associated with a listening task, to predict content and outcomes, or to fill in missing information.
- Elaboration: Using prior knowledge from outside the text or conversational context and relating it to knowledge gained from the text or conversation in order to embellish one's interpretation of the text.
- Prediction: Anticipating the contents and the message of what one is going to hear.
- Contextualization: Placing what is heard in a specific context in order to prepare for listening or assist comprehension.
- Reorganizing: Transferring what one has processed into forms that help understanding, storage, and retrieval.
- Translation: Relying on one's knowledge of the first language or additional languages to make sense of what is heard.
- Managing emotions: Keeping track of one's feelings and not allowing negative ones to influence attitudes and behaviors.

Test-taking strategies

Descriptions are based on A. D. Cohen (2011)

- Test-management strategies: Controlled and goal-directed mental actions to find an answer to a question. Informed by both the test paper (the questions, answer options etc.) and the listening text. E.g. choosing an answer option out of four based on the meaning of the relevant passage.
- Test-wiseness strategies: Controlled and goal-directed mental actions to find an answer to a question, informed solely by the test paper (the questions, answer options etc.) or construct-irrelevant external knowledge. Not informed by the listening text itself. E.g. choosing an answer based only on its position among the other answers (a, b, c or d); guessing.

Anxiety

Descriptions are based on Cassady and Johnson (2002) and Horwitz (2010)

- Listening anxiety: Worries, stress, or concerns related to listening in a foreign language.
- Test-taking anxiety: Worries, stress, or concerns related to evaluative situations.

I coded all documents using the qualitative data analysis software Atlas.ti (version 8.4 for Mac). Initially, I imported all data files into Atlas.ti and programmed the codes according to the coding scheme. The data files were then split into segments (referred to as quotations in Atlas.ti) corresponding to the codes in the coding scheme. For example, the following excerpt from Participant 1 was separated into two quotations:

[Quotation 1] Yes, because then I heard “this is the newest release“ [Quotation 2] and then I chose this answer.

Quotation 1 indicates that the participant put individual words into a syntactic pattern to form the bare meaning of the utterance at clause level, so this quotation was coded as **parsing**. Quotation 2 shows that the participant chose an answer to a question based on their understanding of the recording, thus, this quotation was coded as **test-management**. Another example of segmenting is the following:

[Quotation 1] Here I thought, ok now she will probably soon mention what Lego means for him. [Quotation 2] And then I already focused on that.

Quotation 1 shows that the participant tried to predict what the speaker was going to say next. This segment was therefore coded as **prediction**. Quotation 2 indicates that the participant focused on the listening text in a particular way, so this quotation was coded as **focusing attention**.

Whenever I could not unambiguously assign a single code to a quotation, two codes were applied to the same quotation. This was mostly the case for quotations relating to test-management, as test-management was sometimes only evident in combination with either a cognitive process or a listening strategy. The following excerpt is a typical example:

Then I read through everything again, checked everything, and looked whether I hear it again/whether I would choose the same answer again.

In this quotation the participant checked the outcomes of their listening comprehension against the answers they chose, so this was coded as **evaluation**. However, because the participant referred to the test paper (the answers they chose), the quotation was also coded as **test-management**. The quotation could not have been segmented further to differentiate between the two codes, as both codes applied to the quotation as a whole.

During the process of coding three additional codes emerged based on recurring meta-commentary which appeared to be relevant for answering the research questions. Two of these codes were assigned to quotations which were specifically related to the single play/double play condition:

- **Different behaviour:** Reporting different listening behaviour between the first play and second play or between single play and double play. Example:
So most times during the second play/during the first play one hears the individual things more/so the things one is waiting for the whole time and during the second play one understands more about the context and more about the other details [...].
- **Prefer double play:** Indicating a preference for double play compared to single play. Example:

It is nice when I can hear it twice, because then I can explicitly listen to that again, because I was relatively sure about the other questions, I don't have to pay attention to that any more. But then I can see clearly which ones I haven't answered so that I can answer these. That is nicer in double play.

One of the additional codes was assigned to comments in relation to the research methodology, to discern whether the research methodology was inhibiting natural processing during task completion:

- **Reactivity:** Indicating that the research method is distracting or is inhibiting natural processing. Example:

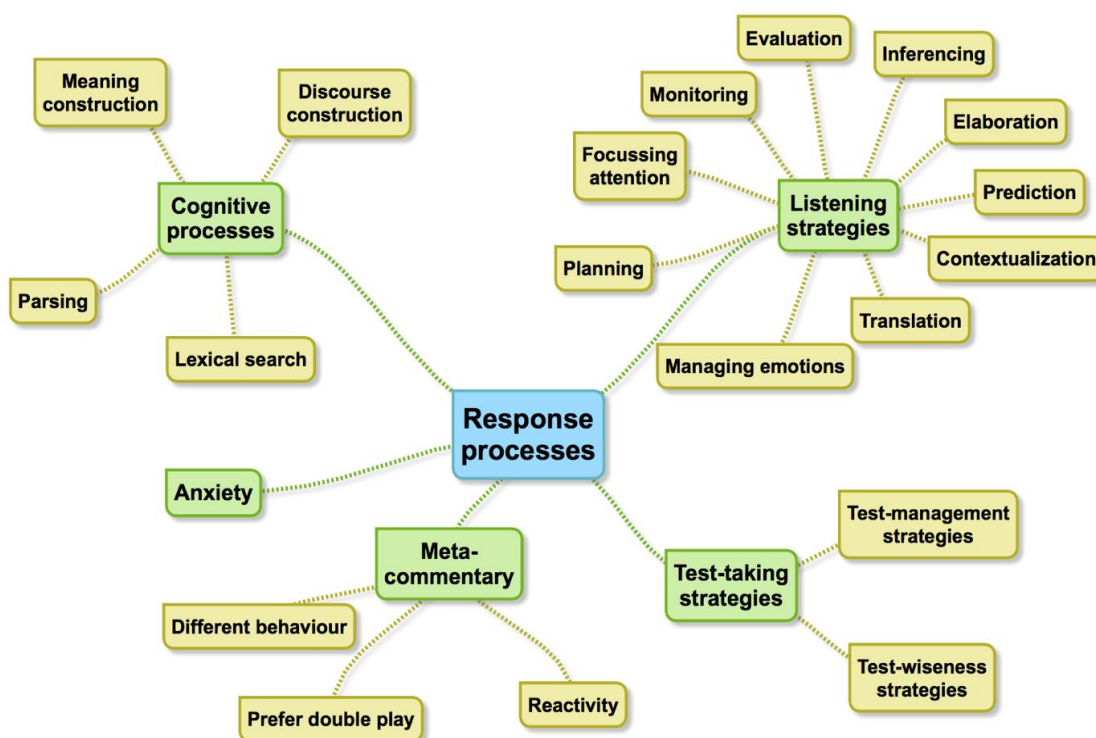
What irritated me a little bit was the thing with the head (referring to the fact that they had to keep their head still for accurate eye-tracking readings), it was not very bad, but I wasn't used to it, only moving my eyes.

No evidence was found in the data for two of the original codes: the cognitive process *input decoding* and the listening strategy *reorganizing*. For input decoding, other studies have also found it difficult to isolate evidence from verbal report data (see, for example, Brunfaut & McCray, 2015; or Holzkecht et al., 2017) or had to rely on evidence from written notes to draw conclusions about mishearings (Rukthong, 2015). However, input decoding underlies all other cognitive processes (Field, 2013) and it is therefore reasonable to assume that test takers who engaged in cognitive processing would have relied on input decoding at a fundamental level, but it was so prevalent and automatised so as not to be retrievable by participants in their reporting behaviour. The strategy of reorganising, on the other hand, is operationalised by processes such as writing a summary, repeating words or phrases out loud, grouping information, or taking notes while listening (Vandergrift & Goh, 2012, p. 282). It is likely that reorganising was not observed because participants did not have the opportunity nor the time to engage in these processes while completing the tasks.

Finally, the two separate codes on anxiety (test-taking anxiety and listening anxiety) were merged after the coding. This was done as participants only indicated general levels of anxiety and did not specify whether their anxiety was related to the test-taking process or to listening in a foreign language.

In summary, the final coding scheme consisted of 20 individual codes grouped into five main categories: cognitive processes, listening strategies, test-taking strategies, anxiety, and meta-commentary. The codes are summarised in Figure 9. For ease of reading, further definitions of the coding categories and illustrative excerpts from the data will be provided in Chapter 5.

Figure 9: Study 2: final coding categories



3.6.5.3. Inter-coder reliability

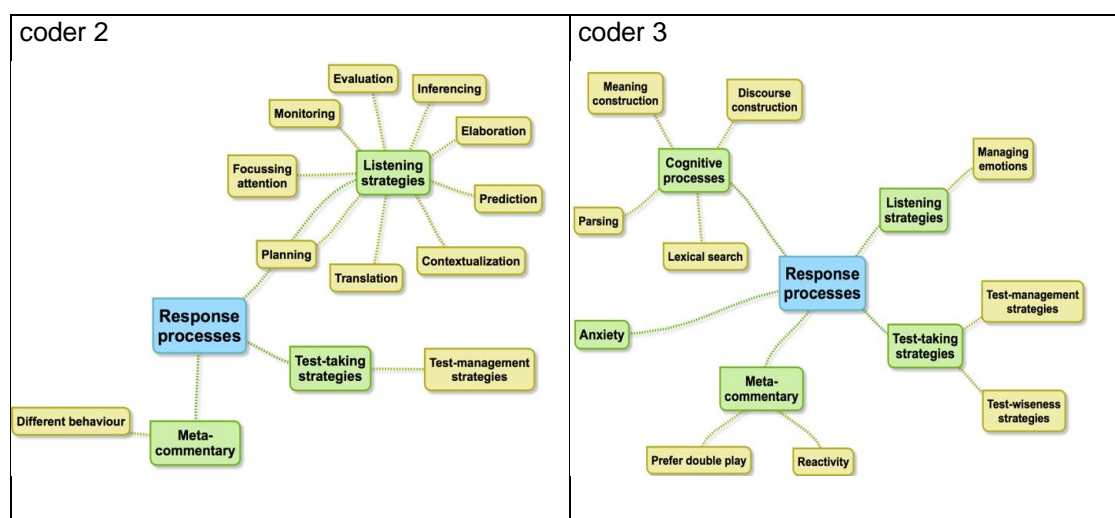
Once all of the data was coded by me, 10 percent of all quotations were double-coded by a second coder (coder 2) and another 10 percent by a third coder (coder 3) to establish reliability of the coding process. The two additional coders were language assessment specialists and University lecturers with many years of experience in listening test development. One of them held an MA in teaching English and an MA in language testing and was in the process of completing a PhD in language testing, while the other held an MA in teaching English, an MA in language testing, and a PhD in applied linguistics with a focus on language testing.

Due to the large number of codes, the two coders focussed on different parts of the coding scheme and were therefore assigned different quotations. Coder 2 was assigned quotations which were initially coded as listening strategies by me. In addition,

the data for coder 2 also included quotations originally coded as test-management, as test-management strategies were sometimes evident in combination with a listening strategy (see discussion in Section 3.6.5.2 above), as well as the meta-commentary “different behaviour”. Coder 3, on the other hand, was assigned quotations originally coded as cognitive processes, test-taking strategies (including test-management and test-wiseness), anxiety, and the remaining meta-commentary codes which emerged during the coding (“reactivity” and “prefer double play”). In total, coder 2 focussed on 11 codes and coder 3 on 10 codes, as illustrated in Figure 10.

To get a representative sample of the data for the double-coding process, I made sure to include quotations from all participants, all tasks, both conditions, and all stages of recall for both coders. Each coder was sent their version of the coding scheme, including short descriptions of the codes and an example from the data for each code, and a coding document, which included the quotations to be coded and a column to enter the codes (see Appendix 7). They were asked to familiarise themselves with the coding scheme first and then assign a code to each of the quotations in the coding document.

Figure 10: Study 2: coding categories for double-coding for coder 2 and coder 3



Following the double-coding, I calculated inter-coder agreement between myself and the double-coders. To that end, I first transformed my original codings as well as the double-coders’ codings into nominal data by assigning 0 (code not applied) and 1 (code applied) for each code and each quotation. The data for all quotations were then entered in Microsoft Excel and Gwet’s AC₂ was calculated separately for each double-

coder, using the Excel add-in by Zaiontz (2019). Gwet's AC₂ was chosen as it is considered more robust than other inter-coder reliability coefficients such as Cohen's kappa, Fleiss's kappa, Conger's kappa, or Krippendorff's alpha (Quarfoot & Levine, 2016).

As a final step, the extent of agreement was calculated according to the benchmarking procedure suggested by Gwet (2014, pp. 164–181). Instead of simply using the reliability coefficient for interpreting the strength of agreement between coders, which might mask the true level of agreement as the associated error of measurement is not taken into account, Gwet suggests to utilise the standard error to calculate the coefficient's membership probabilities for each range on a given benchmark scale. To classify the extent of agreement, the benchmark scale by Landis and Koch (1977) was used, which differentiates between poor (<0.00), slight (0.00-0.21), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), and almost perfect (0.81-1.00) agreement. The membership probabilities for each of these categories are reported.

The results are displayed in Table 10 and Table 11. The tables include Gwet's AC₂ for each coding category and for the overall agreement with each coder, as well as the associated standard errors and the membership probabilities (in percent) in relation to Landis and Koch's (1977) benchmark scale. The highest membership probability is highlighted in each row to make the results more immediately interpretable. As shown in the tables, inter-coder agreement was high. The overall agreement between coder 2 and myself was almost perfect (with a 93 percent probability) and between coder 3 and myself it was substantial (with a 62 percent probability) to almost perfect (38 percent probability). Agreement was also calculated separately for the four main response processes of interest to inspect whether certain code groups attracted more agreement than others. As shown in Table 10, for listening strategies agreement between coder 2 and myself was substantial (74 percent probability) to almost perfect (26 percent probability) and for test-taking strategies (only test-management) it was almost perfect (99 percent probability). For coder 3 (Table 11), agreement for cognitive processes was substantial (83 percent probability) to almost perfect (15 percent probability) and for test-taking strategies (test-management and test-wiseness) it was closer to almost perfect (84 percent probability) than substantial (16 percent probability). The agreement probabilities for anxiety are more scattered due to there being only one coding category and the small number of quotations which received that code in the double-coding

document (N=7), which resulted in a high standard error. Coder 3 and myself agreed in 6 out of 7 cases that a quotation was related to anxiety, resulting in substantial (34 percent probability) to almost perfect (57 percent probability) agreement. Despite these high levels of inter-coder agreement, I discussed all quotations where there was disagreement with the two double-coders to reach a consensus decision for each case. I then individually double-checked all of my original codings in light of the discussions.

Table 10: Study 2: inter-coder agreement between the researcher and coder 2

Response processes	Gwet's AC ₂	S.e.	Probability (in percent) for agreement to be		
			moderate	substantial	almost perfect
listening strategies	0.765	0.054	0	74	26
test-taking strategies	0.935	0.054	0	1	99
overall	0.864	0.035	0	7	93

Table 11: Study 2: inter-coder agreement between the researcher and coder 3

Response processes	Gwet's AC ₂	S.e.	Probability (in percent) for agreement to be		
			moderate	substantial	almost perfect
cognitive processes	0.733	0.063	2	83	15
test-taking strategies	0.822	0.071	0	16	84
anxiety	0.831	0.171	9	34	57
overall	0.774	0.045	0	62	38

4. Results Study 1

The results of Study 1 are presented in six main sections. First, the findings of a CTT analysis are outlined in Section 4.1, including the items' facility values as well as discrimination and reliability indexes across the two conditions (single play and double play). In Section 4.2 the results of an IRT analysis utilizing MFRM are described. MFRM was used to explore to what extent, in terms of average item difficulty, the different tasks and task types were impacted by the listening condition (single play or double play). Next, a detailed analysis of candidates' answer changes in the double play condition is presented in Section 4.3, based on the candidates' use of different coloured pens in the first and second play of the double play condition. The chapter then goes on to describe the results of Questionnaire 1, which targeted candidates' strategic behaviour and anxiety levels, in Section 4.4. The questionnaire data was analysed by means of descriptive statistics, an exploratory factor analysis, and a Wilcoxon signed-rank test to explore statistically significant differences between single play and double play. Section 4.5 then outlines the findings of Questionnaire 2, which included questions on the candidates' topic and task familiarity, their perceived task difficulty and perceived validity, as well as their preference for single or double play. Finally, a summary of the main findings is presented in Section 4.6.

4.1. Classical test theory

CTT was used to calculate test and item parameters (facility values, discrimination indexes, and reliability indexes) to inform RQ 1 and RQ 1a. As outlined in section 3.5.5, due to the research design the 306 participants were divided into two sub-groups. Participants from sub-group 1 took the MC tasks in a double play condition and the NF tasks in a single play condition and participants from sub-group 2 took the MC tasks in a single play condition and the NF tasks in a double play condition. I chose to analyse each sub-group separately as calculating the discrimination required that the test is treated as a whole for the purposes of CTT.

Table 12 displays the results for the MC tasks in a double play condition and the NF tasks in a single play condition (sub-group 1). Cronbach's Alpha was 0.819, so the overall reliability of the test was high (Pallant, 2007, p. 98). The following are highlighted in orange as indicators of the items not performing as expected:

discrimination indexes (corrected item-total correlation) of 0.25 or less (see Henning, 1987), items which when deleted would increase the overall reliability (Cronbach's Alpha if item deleted), and facility values below 0.20 and above 0.80 if the item had a low discrimination index or impacted Cronbach's Alpha negatively (see Bachman, 2004, p. 138). As can be seen in the table, half of the MC items had low item discrimination and one of them, if deleted, would increase the overall reliability of the test (item 6 in MC 1). The NF tasks performed better, with only 4 out of 18 items displaying low item discrimination and one of them increasing the overall reliability if deleted (item 6 in NF 2)

Table 12: Study 1: reliability, facility values, and discrimination indexes for the MC tasks in a double play condition and the NF tasks in a single play condition (sub-group 1)

Cronbach's Alpha: 0.819				
Item	Facility value	Corrected item-total correlation	Cronbach's Alpha if item deleted	N
MC 1 double q1	0.73	0.244	0.817	153
MC 1 double q2	0.63	0.222	0.818	153
MC 1 double q3	0.80	0.293	0.815	153
MC 1 double q4	0.91	0.263	0.816	153
MC 1 double q5	0.50	0.325	0.814	153
MC 1 double q6	0.97	0.041	0.820	153
MC 2 double q1	0.81	0.171	0.819	153
MC 2 double q2	0.61	0.232	0.818	153
MC 2 double q3	0.67	0.515	0.807	153
MC 2 double q4	0.76	0.294	0.815	153
MC 2 double q5	0.82	0.218	0.818	153
MC 2 double q6	0.65	0.434	0.810	153
NF 1 single q1	0.54	0.439	0.809	153
NF 1 single q2	0.81	0.317	0.814	153
NF 1 single q3	0.69	0.568	0.805	153
NF 1 single q4	0.63	0.244	0.817	153
NF 1 single q5	0.63	0.497	0.807	153
NF 1 single q6	0.65	0.428	0.810	153
NF 1 single q7	0.32	0.453	0.809	153
NF 1 single q8	0.18	0.297	0.815	153
NF 1 single q9	0.73	0.335	0.814	153
NF 2 single q1	0.25	0.263	0.816	153
NF 2 single q2	0.20	0.308	0.815	153
NF 2 single q3	0.31	0.238	0.817	153
NF 2 single q4	0.65	0.388	0.812	153
NF 2 single q5	0.38	0.492	0.807	153
NF 2 single q6	0.43	0.160	0.821	153
NF 2 single q7	0.43	0.348	0.813	153
NF 2 single q8	0.27	0.453	0.809	153
NF 2 single q9	0.33	0.207	0.819	153

The results for the MC tasks in a single play condition and the NF tasks in a double play condition (sub-group 2) are presented in Table 13. Cronbach's Alpha was again high with 0.833. As in the test for sub-group 1, half of the MC items had low item discrimination, but this time two of them would increase the overall reliability if deleted (item 3 and 6 in MC 1). The statistics for the NF tasks are considerably better, with only 2 out of 18 displaying low item discrimination and one of them increasing the overall reliability if omitted (item 9 in NF 2).

Table 13: Study 1: reliability, facility values, and discrimination indexes for the MC tasks in a single play condition and the NF tasks in a double play condition (sub-group 2)

Cronbach's Alpha: 0.833				
Item	Facility value	Corrected item-total correlation	Cronbach's Alpha if item deleted	N
MC 1 single q1	0.74	0.314	0.829	153
MC 1 single q2	0.63	0.380	0.826	153
MC 1 single q3	0.86	0.110	0.834	153
MC 1 single q4	0.92	0.207	0.831	153
MC 1 single q5	0.42	0.276	0.831	153
MC 1 single q6	0.93	0.047	0.835	153
MC 2 single q1	0.74	0.323	0.828	153
MC 2 single q2	0.71	0.378	0.826	153
MC 2 single q3	0.71	0.231	0.832	153
MC 2 single q4	0.76	0.223	0.832	153
MC 2 single q5	0.80	0.239	0.831	153
MC 2 single q6	0.65	0.340	0.828	153
NF 1 double q1	0.75	0.454	0.824	153
NF 1 double q2	0.91	0.154	0.833	153
NF 1 double q3	0.86	0.432	0.825	153
NF 1 double q4	0.92	0.313	0.829	153
NF 1 double q5	0.76	0.396	0.826	153
NF 1 double q6	0.78	0.333	0.828	153
NF 1 double q7	0.64	0.613	0.817	153
NF 1 double q8	0.45	0.570	0.819	153
NF 1 double q9	0.84	0.378	0.827	153
NF 2 double q1	0.52	0.357	0.827	153
NF 2 double q2	0.49	0.356	0.827	153
NF 2 double q3	0.70	0.492	0.822	153
NF 2 double q4	0.94	0.263	0.830	153
NF 2 double q5	0.74	0.533	0.821	153
NF 2 double q6	0.58	0.318	0.829	153
NF 2 double q7	0.88	0.418	0.826	153
NF 2 double q8	0.56	0.554	0.819	153
NF 2 double q9	0.71	0.165	0.834	153

When comparing the same task types across the two conditions, it can be seen that item statistics are generally slightly better in the double play condition. For the MC

tasks, although six items displayed low discrimination indexes in both the single play and the double play condition, the mean discrimination index of the six items was lower in the single play condition (0.175 compared to 0.188). In addition, two items had a negative impact on the overall reliability in the single play condition compared to only one item in the double play condition. Overall, the number of highlighted cells for the MC tasks is higher in the single play condition (12) than in the double play condition (10). The tendency is the same for the NF tasks. In terms of item discrimination, four items had a corrected item-total correlation of 0.25 or less in the single play condition, compared to only two items in the double play condition. The number of items that had a negative impact on the overall reliability was the same between the two conditions (one in each), but overall the number of highlighted cells is again higher in the single play condition (5) than in the double play condition (4).

Another interesting finding emerges when comparing the mean facility values for each task across the two sub-groups. As shown in Table 14 and Table 15, the tasks' mean facility values were impacted differently by the listening condition. For the MC tasks the mean facility values are very similar between the two sub-groups, although it might have been expected that they would be lower for sub-group 2 as they completed the tasks in a single play condition. For the NF tasks, sub-group 2, who completed the tasks in a double play condition, clearly outperformed sub-group 1, who completed the tasks in a single play condition. While it could be the case that some property of MC tasks make them less susceptible to the beneficial effects of double play, another potential explanation for this finding is that sub-group 2 may have been more proficient than sub-group 1. However, this initial CTT analysis, which was used mainly to explore the patterns in the data, was not informative in this regard (see also the discussion in Section 3.2). Therefore, in order to test this further, an IRT analysis was performed, as outlined in the following section.

Table 14: Study 1: mean facility values of the MC tasks

Subgroup	Task	Condition	Mean FV
1	MC 1	double play	0.76
	MC 2	double play	0.72
2	MC 1	single play	0.75
	MC 2	single play	0.73

Table 15: Study 1: mean facility values of the NF tasks

Subgroup	Task	Condition	Mean FV
1	NF 1	single play	0.58
	NF 2	single play	0.36
2	NF 1	double play	0.76
	NF 2	double play	0.68

4.2. Many-facet Rasch measurement

MFRM (Linacre, 1994), implemented by the computer program Facets (version 3.81.2), was used to explore to what extent, in terms of average item difficulty, the different tasks and task types were impacted by the listening condition (single play or double play). MFRM models are an extension of the basic Rasch model, which expresses the difficulty of items and the ability of test takers as a probabilistic function on a latent variable (for details see Eckes, 2015, pp. 21–27). In contrast to the basic Rasch model, MFRM has the advantage that multiple variables (or facets), such as test takers, items, tasks, task types, or listening conditions, can be modelled jointly without the need to average across variables (Eckes, 2015; McNamara, 1996). This was necessary because a between-participants research design was used – the two task formats were completed in different conditions by two sub-groups of participants – and the two sub-groups may have differed in their average listening proficiency.

For each Facets analysis a specifications file needs to be created. The specifications file includes information on the number of facets included in the analysis, the elements for each of the facets, and the model according to which the data is analysed. One of the three specification files used for the analysis of the data is included in Appendix 9. It includes five facets: students (306 elements corresponding to the 306 students; please note that the students are not listed in the appendix), task types (two elements: MC and NF), tasks (four elements: two tasks for each task type), conditions (two elements: single play and double play in the appendix), and items (30 elements corresponding to the 30 items; only the first two items are shown in the appendix). At the bottom of the specifications file the data is inserted (the data is not shown in the appendix).

The model statement in Appendix 9 is *Model=?,?B,?B,?B,?,D*. It specifies that all elements of all facets interact with each other, indicated by the question marks for each facet. The “B” for facets 2, 3 and 4 indicate that a bias/interaction analysis is performed including these three facets (see Eckes, 2015, pp. 133–139). The bias analysis was used to inform RQ 1. Finally, the “D” specifies that the data was scored dichotomously (correct or incorrect response for each item).

Three of the facets were defined as dummy facets, which means that their measure was anchored at 0 logits for the analysis. This was done for task types and tasks in order to avoid disjoint subsets in the data, as the items were nested within task types and tasks.

In order to still get average item difficulty for the task types and tasks the items were grouped according to the four tasks. The conditions facet was also defined as a dummy facet to be able to perform a bias analysis with task types and tasks (see Linacre, 2019).

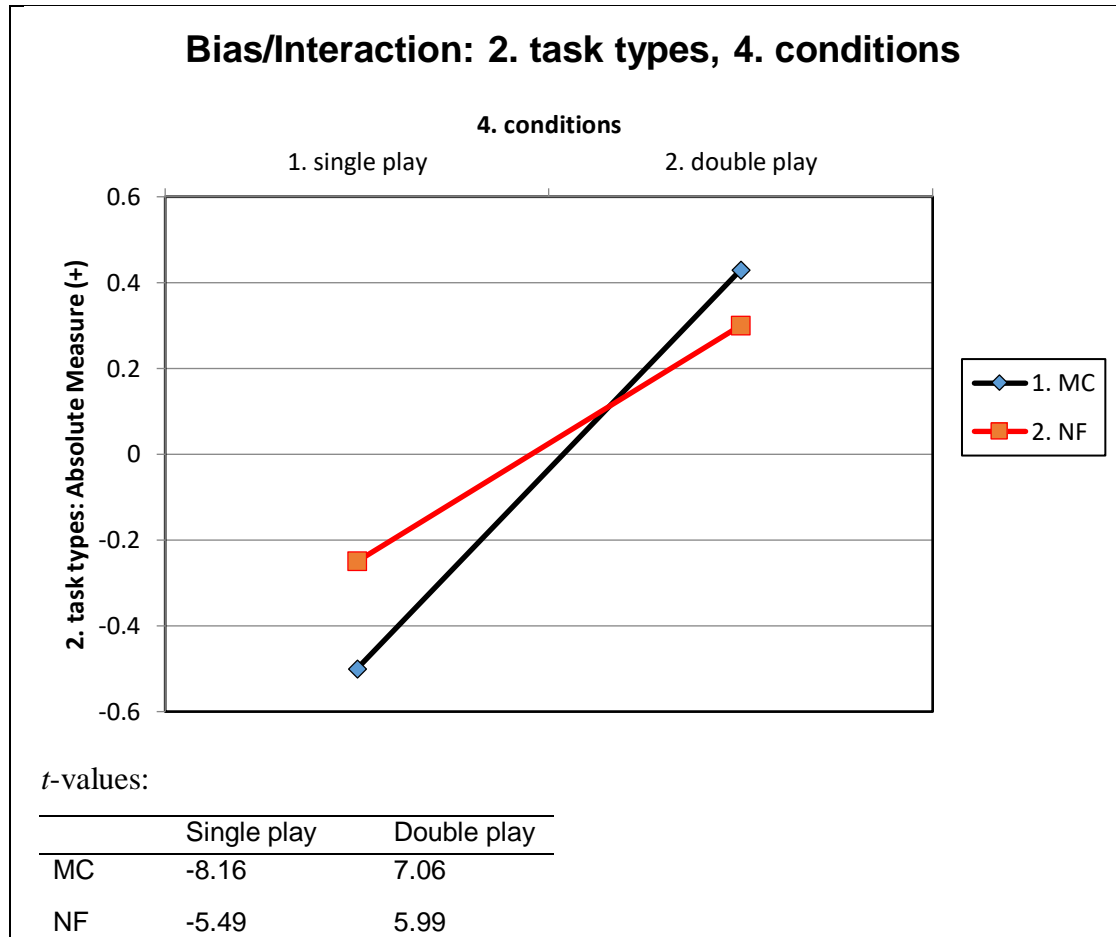
The analysis was run three times across all participants to answer RQ 1. First, the single play condition was compared to the double play condition in a bias analysis to inspect potentially significant differences in average item difficulty between the two conditions. This was done to compare the findings with the main share of previous research in this area, which has found that double play generally makes test items easier. In a second bias analysis, the first play in the double play condition was compared to the single play condition, again to investigate differences in test scores, as this has not been studied to date. Although Field (2015) initially set out to investigate this, he decided not to pursue it further as “[i]n the event [...] it proved hard to conclusively identify marked differences of behaviour during a single play as compared with the first hearing of a double play” (Field, 2015, p. 12). Finally, the first play and the second play of the double play condition were compared in a third bias analysis, to establish the difference in average item difficulty between these two conditions. As outlined above, the three bias analyses were run twice – once relating to the two task types and once relating to the four individual tasks – to investigate whether the conditions impacted the task types and tasks differentially. The Facets specifications file was the same for all three analyses, except for facet 4 (as the conditions were different each time). The results for each of the three analysis will be outlined below.

4.2.1. Single play versus double play

The first analysis investigated the interaction between the listening condition (single play and double play) and the two task types as well as the four tasks. The Facets bias analysis showed that the average item difficulty of the two task types is impacted similarly by the single and double play condition (Figure 11). As shown in previous research, the items get easier in the double play condition. However, there was also a difference between the two task types. Overall, the MC tasks are more difficult than the NF tasks in the single play condition, but easier than the NF tasks in the double play condition. The difference in average item difficulty between single play and double play was .93 logits for MC tasks and .55 logits for NF tasks. All associated *t*-values in Figure

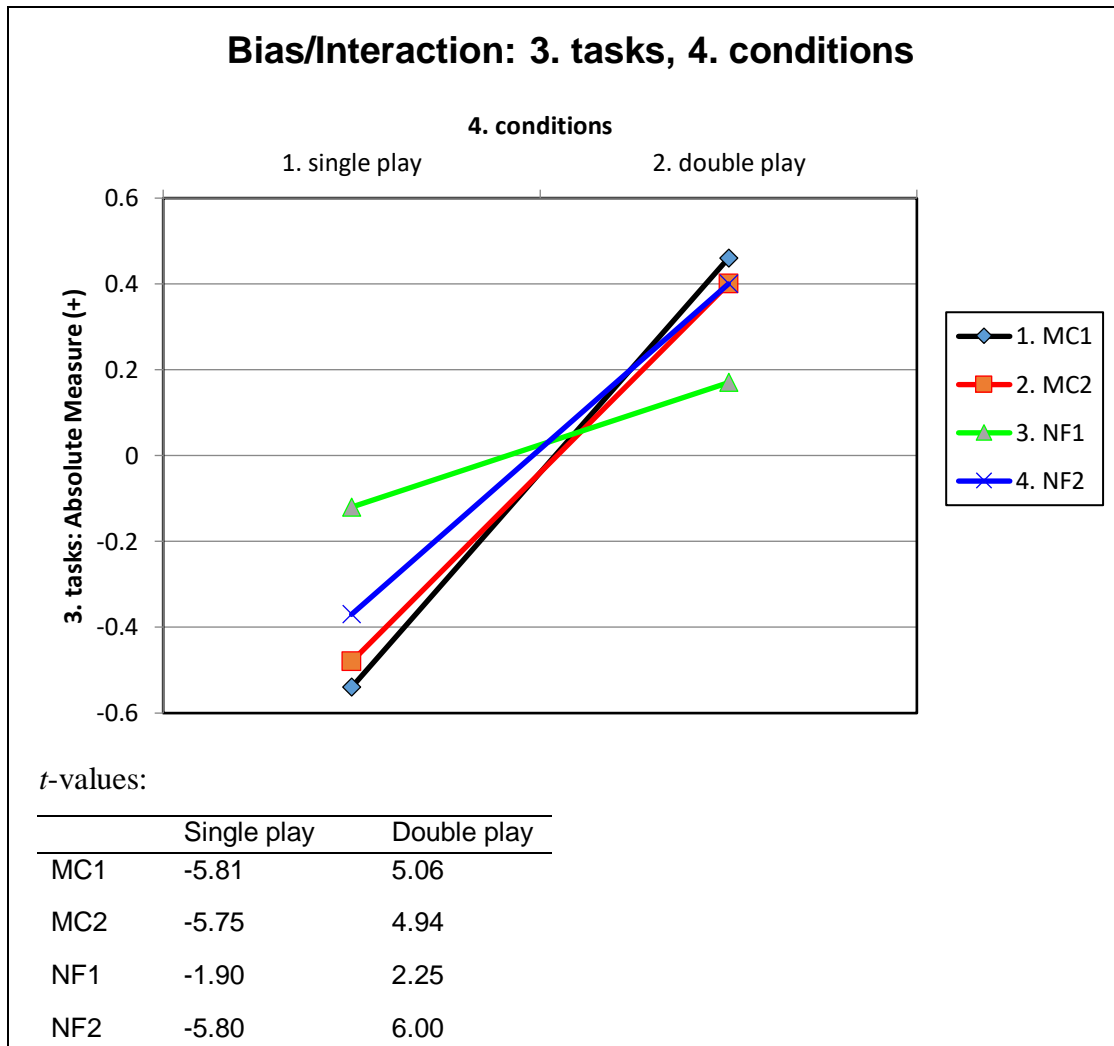
11 are larger than ± 2.00 , which indicates that the bias in difficulty is significant (McNamara, 1996; McNamara, Knoch, & Fan, 2019, pp. 122–124).

Figure 11: Study 1: Facets bias analysis and associated t -values between single play and double play across the two task types



This finding is confirmed by the bias analysis of the four individual tasks (Figure 12), which showed that the MC tasks' difficulty is impacted more than the NF tasks' difficulty, particularly NF1. It can also be seen that in the double play condition the difference in task difficulty between the four tasks is smaller than in the single play condition. In single play, task difficulties ranged from $-.12$ logits to $-.54$ logits (a range of $.42$ logits), whereas in double play task difficulties spanned between $.17$ and $.46$ logits (a range of $.29$ logits). The t -values of all but one possible pairs in Figure 12 are larger than ± 2.00 , which again indicates that the bias in difficulty is significant. Only for NF 1 the lower t -value was -1.9 .

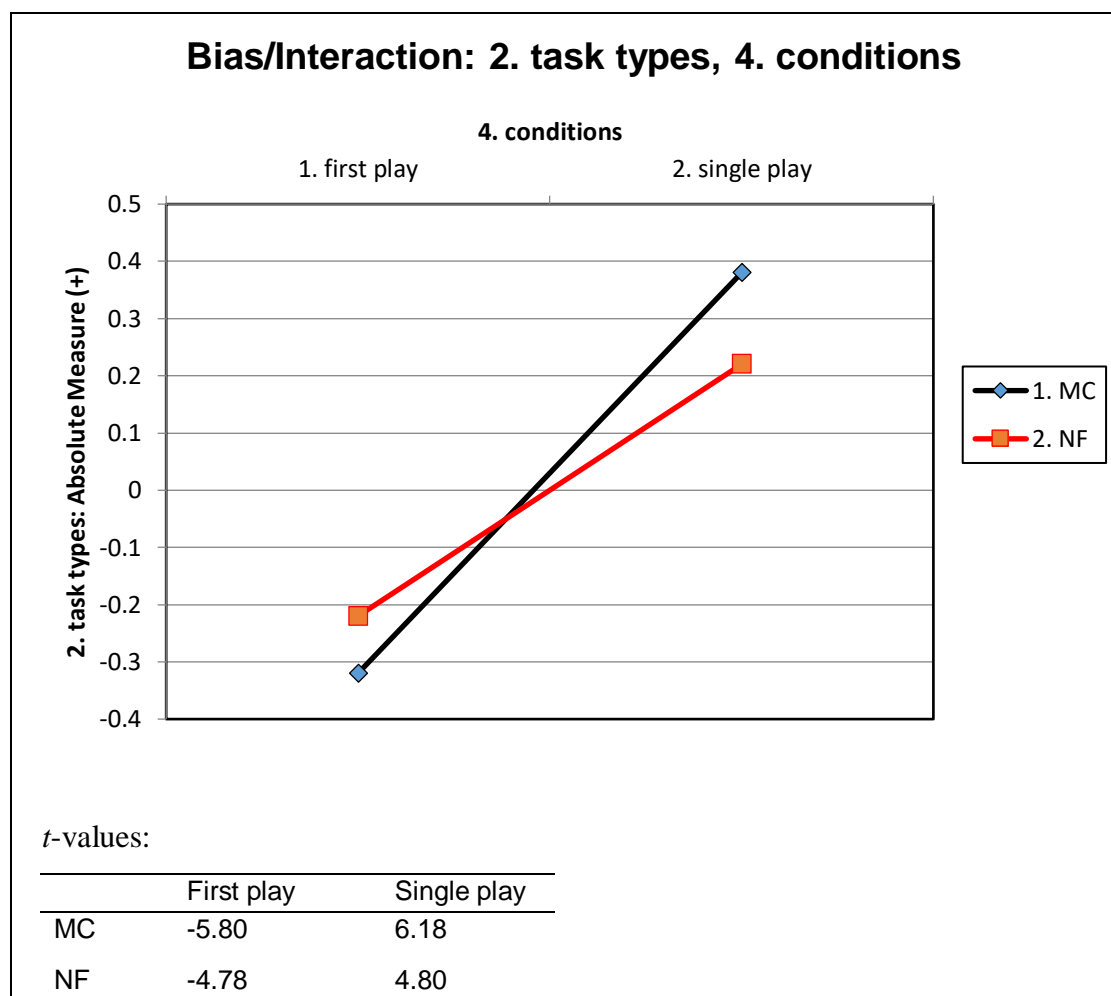
Figure 12: Study 1: Facets bias analysis and associated t-values between single play and double play across the four tasks



4.2.2. First play versus single play

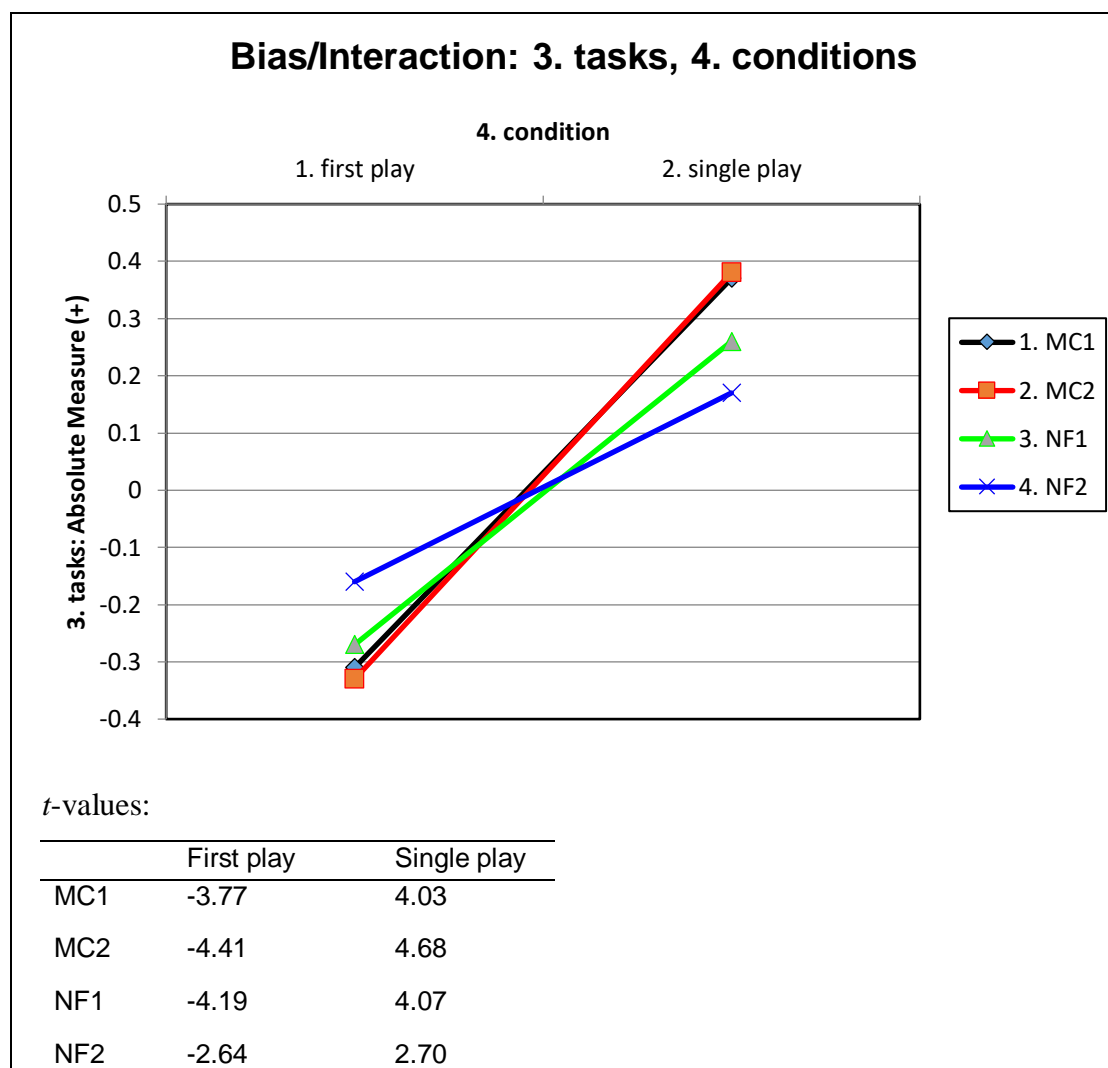
A second bias analysis was conducted to detect potential differences in average item difficulty between the first play of the double play condition and the single play condition, as previous research reports that it was difficult to detect differences in test taker behaviour between these two conditions (Field, 2015, p. 12). As discussed in section 3.5.6, the scores for the first play of double play were derived from the answers marked in blue. Whenever participants had not selected an answer for a question yet, this was marked as zero for the question.

Figure 13: Study 1: Facets bias analysis and associated t -values between the first play in double play and single play across the two task types



The results show that there is a significant difference in average item difficulty between the two conditions for the two task types (Figure 13) and also for all four individual tasks (Figure 14). The t -values for all possible pairs exceeded ± 2.00 and thus indicate statistical significance. As can be seen, students scored higher across all tasks in the single play condition than in the first play of the double condition. Again, the bias is larger for the MC tasks than the NF tasks. The difference in average item difficulty between single play and the first play of double play was .68 logits for MC1, .71 logits for MC2, .53 logits for NF1, and .31 logits for NF2.

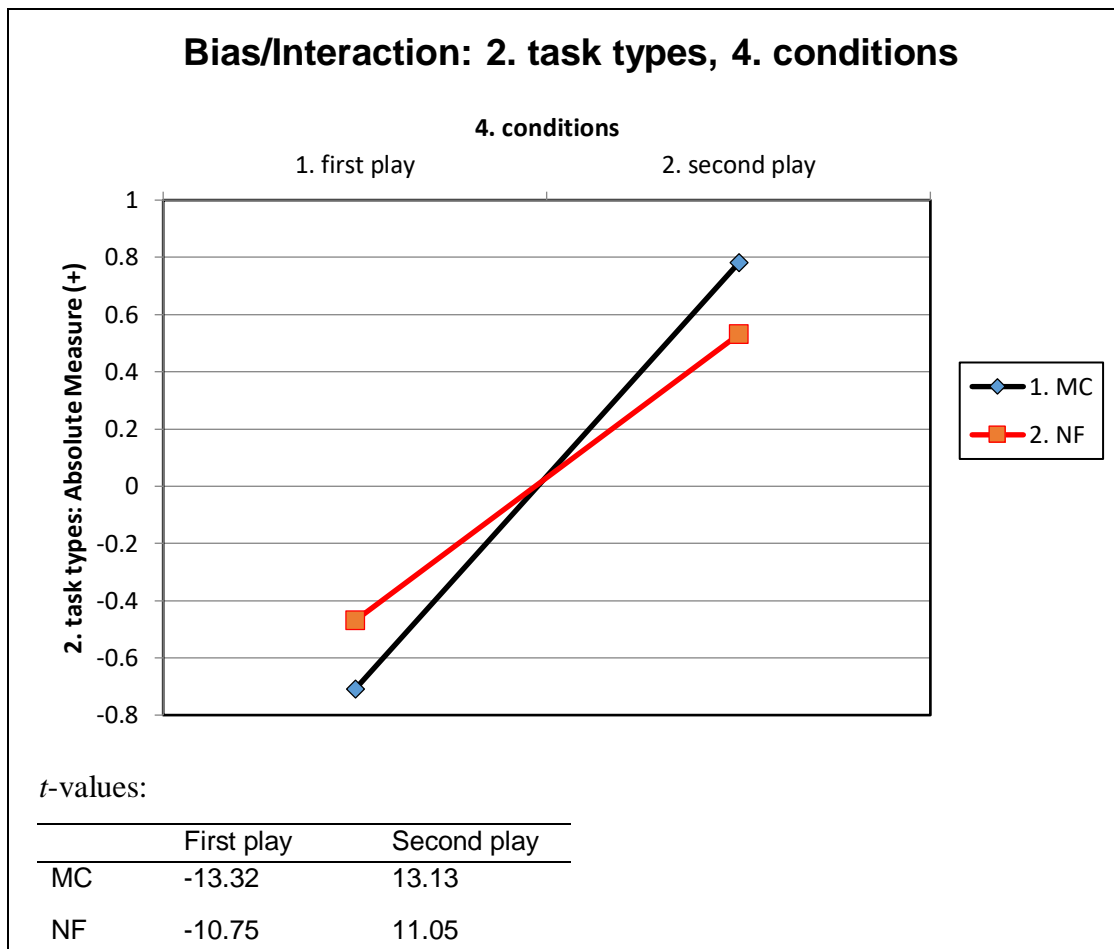
Figure 14: Study 1: Facets bias analysis and associated t-values between the first play in double play and single play across the four tasks



4.2.3. First play versus second play

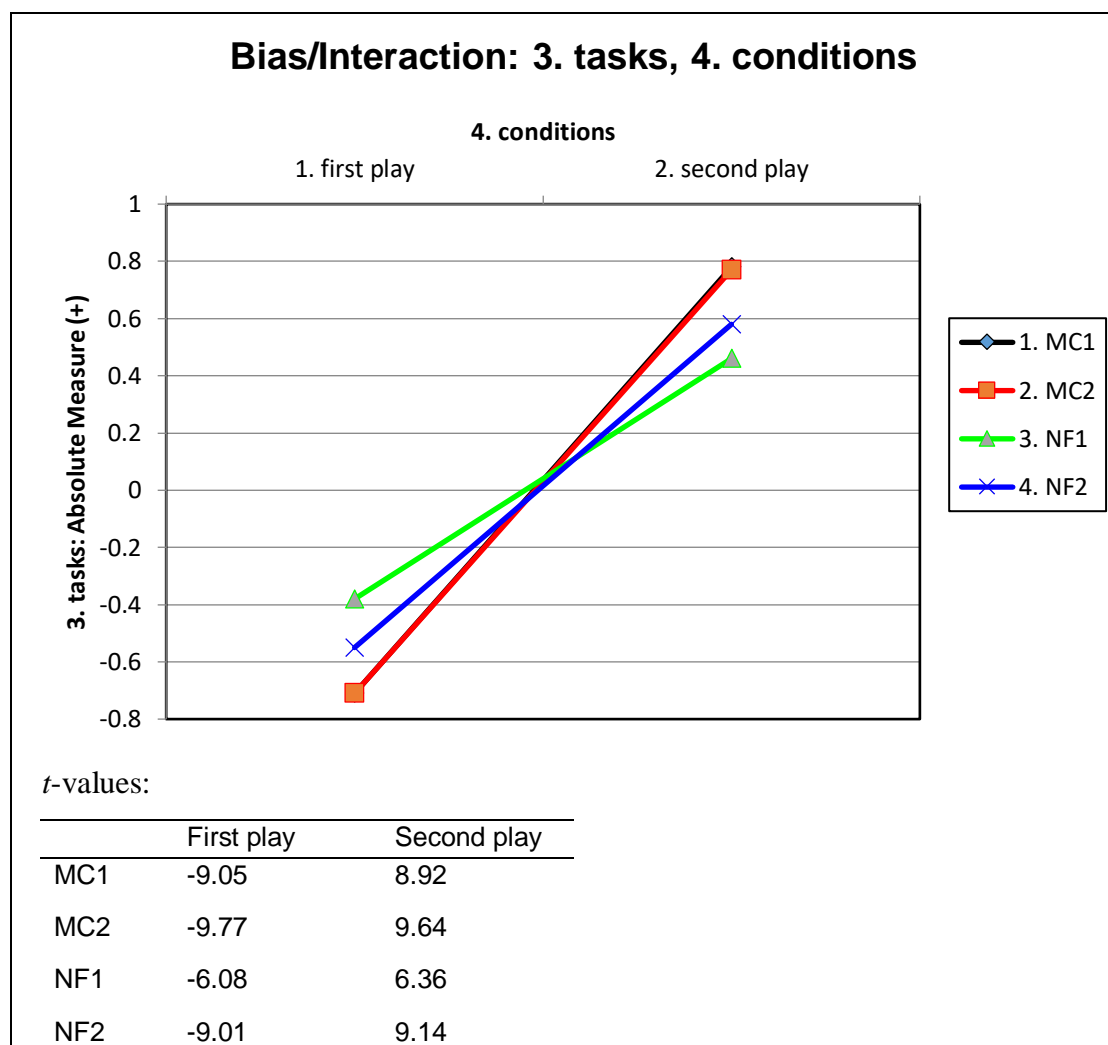
A third bias analysis was conducted for the first and the second play in the double play condition across both task types and all tasks. As described in Section 3.5.6, the items completed in the double play condition were scored twice according to the keys and the extended marking schemes: once for answers after the first listening (as indicated by the blue pen) and once for answers after the second listening (as indicated by the red pen). The analysis compared the answers marked in blue with the answers marked in red.

Figure 15: Study 1: Facets bias analysis and associated t -values between the first play and the second play in double play across the two task types



As expected, average item difficulty was significantly higher after the first play than the second play. In terms of task type effects, the same pattern can be observed, in that students benefitted more from the second play in MC tasks than in NF tasks (Figure 15 and Figure 16). Bias was again significant for all possible pairs, with a difference in average item difficulty between single play and double play of 1.49 logits for MC1, 1.48 logits for MC2, .84 logits for NF1, and 1.13 logits for NF2. All associated t -values were larger than -6 and 6 and thus indicate statistical significance.

Figure 16: Study 1: Facets bias analysis and associated t-values between the first play and the second play in double play across the four tasks

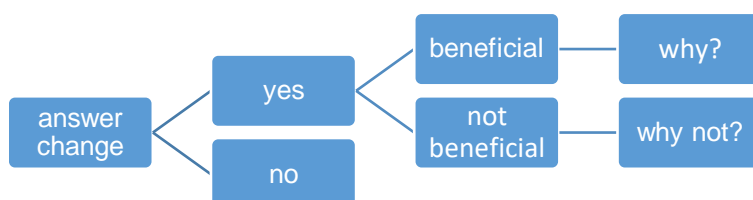


In summary, the MFRM analysis showed that items were easier in double play compared to single play, but that this was less pronounced for the NF tasks. Particularly, NF1 received the least amount of benefit in the double play condition. However, the findings also indicate that listeners were behaving differently in the first play of double play versus single play, as in the first play of double play average item estimates were significantly lower than in single play. This indicates that double play is not simply a repetition of a single play condition, but rather that test takers are behaving differently from the beginning.

4.3. Answer change during the second play

As outlined in Section 3.5.5, students used a blue pen during the first play and a red pen during the second play for the tasks completed in the double play condition to be able to analyse in what way they changed their answers during the second play. It was hoped that this analysis would inform RQ 3 on listening strategies and RQ 4 on test-taking strategies. The analysis was run separately for each task type according to 1) whether there was an answer change during the second play or not; 2) where there was a change, whether the change was beneficial or not; 3) where the change was beneficial, in what way it was beneficial; and 4) where the change was not beneficial, why it was not beneficial. The analysis process is illustrated in Figure 17.

Figure 17: Study 1: analysis process for answer changes during the second play



Prior to the analysis, students' answers during the second play, as indicated by the use of the red pen in the test booklets, were categorised according to the coding scheme displayed in Table 16. As shown in the table, there were a total of 10 coding categories, with 7 categories for answers that were changed during the second play and 3 categories for answers that were not changed. The answers that were changed were differentiated further according to whether the change was beneficial or not, with 4 distinct categories for a beneficial change and 3 categories for changes that were not beneficial.

Table 16: Study 1: categorization of answer changes during the second play

Change or no change	Beneficial or not beneficial	Category
change	beneficial	0 no answer after first, correct after second
		1 incorrect after first, correct after second
		2 correct after first, more details after second and "more correct" (NF only)
		3 correct after first, different correct answer after second (NF only)
	not beneficial	4 no answer after first, incorrect after second
		5 incorrect after first, changed after second but still incorrect
		6 correct after first, incorrect after second
	no change	7 no answer in either first or second
		8 correct after first, no changes after second
		9 incorrect after first, no changes after second

Two of the categories for a beneficial change were only applicable to NF tasks (categories 2 and 3 in Table 16). Although these two types of changes did not lead to an increase in test scores, they were still regarded as beneficial as students were able to show their listening proficiency to a fuller extent during the second play. For the MC tasks, students may also have understood more details during the second play on a specific question, but they were not able to show that due to the restricted test format. Thus, for the comparisons between the two task types described in the following, categories 2 and 3 for the NF tasks were added to category 8, as otherwise the number of changes would have been artificially inflated for the NF tasks. The two categories were instead analysed separately, and these findings will be presented at the end of the section.

4.3.1. Frequencies of answer change

Table 17 displays the number of times students changed an answer during the second play. As described above, the two separate categories for the NF tasks (categories 2 and 3 in Table 16) were categorised as “no change” for this analysis to be able to compare the two task types. Out of a total of 1,836 answers on the two MC tasks (153 candidates from sub-group 1 times 12 items), 40.6 percent were changed during the second play. For the NF tasks the number of changes during the second play was lower, with 35.5 percent out of 2,754 answers (153 candidates from sub-group 2 times 18 items). There was a statistically significant difference in the number of times students changed their answer during the second play between the two task types: Students who took the MC tasks in a double play condition changed their answers significantly more often than

students who took the NF tasks in the double play condition ($\chi^2(1, N = 1,725) = 4.686$, $p > .03$), with a medium effect size (Cohen's $h = 0.27$). This confirms the findings of the MFRM analysis in section 4.2 above, i.e. that MC tasks seem to be aided more from double play than NF tasks.

Table 17: Study 1: frequencies of answer change during the second play across task types

	MC		NF	
	<i>N</i>	%	<i>N</i>	%
change	746	40.6	979	35.5
no change	1,090	59.4	1,767	64.2
unclear*			8	0.3
total	1,836	100	2,754	100

*One candidate used a white-out pen for eight items on one of the NF tasks, so it was not clear whether she changed her answers during the second play.

4.3.2. Effects of answer change

It was also investigated whether students benefitted from changing their answers during the second play as well as the typology of changes and frequencies of those types. The analysis was run according to the categories outlined in Table 16 above. For the analysis the two NF specific categories in Table 16 (categories 2 and 3) were added to NF category 8 to be able to compare the two task types. The results are displayed in Table 18. The table includes percentages for each category and task type, Chi-square statistics to explore potentially significant differences between the two task types, and effect sizes (Cohen's h) for categories with statistically significant differences. The Chi-square statistics were calculated separately for each category based on a comparison of the proportions for each task type, using the website medcalc.org (Schoonjans, 2018), which utilises the Chi-squared test recommended by Campbell (2007) and Richardson (2011) and the confidence interval calculation recommended by Altman, Machin, Bryant, and Gardner (2000). Effect sizes were calculated manually in SPSS (version 24 for Mac).

Starting with beneficial answer changes at the top of the table, it can be seen that 20.3 percent of all MC answers were left blank during the first play but were correct after the second play (category 0), compared to 17.9 percent of NF answers. There was also a difference in frequency between the two task types for answers which were incorrect after the first play but correct after the second (category 1), with 5.7 percent for the MC tasks and 3.6 percent for the NF tasks. As shown in the table, the differences

between the two task types are statistically significant for both of these categories, with small (category 1) to medium (category 0) effect sizes (see J. Cohen, 1988). When comparing the subtotals for beneficial answer changes, it can be seen that students benefitted significantly more from the second play for MC tasks (25.9 percent of all answers were changed beneficially) than for NF tasks (21.5 percent of all answers were changed beneficially) ($\chi^2(1, N = 1,068) = 11.93, p > .00$), with a small to medium effect size (Cohen's $h = 0.22$). This confirms the findings presented in Section 4.2, where it was shown that students benefitted more from a second play in MC tasks than NF tasks in terms of average item difficulty.

Table 18: Study 1: frequencies and chi-square statistics for the answer change categories across the two task types

Change or no change	Benef. or not benef.	Category	MC %	NF %	Diff. %	95% CI	χ^2	DF	p	Cohen's h
change	benef.	0	20.3	17.9	2.4	0.1 to 4.8	4.15	1	0.04*	0.28
		1	5.7	3.6	2.1	0.9 to 3.4	11.45	1	0.00***	0.05
		<i>subtotal</i>	<i>25.9</i>	<i>21.5</i>	<i>4.4</i>	<i>1.9 to 6.9</i>	<i>11.93</i>	<i>1</i>	<i>0.00***</i>	<i>0.22</i>
	not benef.	4	11.3	10.3	1.0	-0.8 to 2.9	1.15	1	0.28	
		5	1.5	2.6	1.1	0.3 to 1.9	6.31	1	0.01*	0.91
		6	1.9	1.1	0.8	0.1 to 1.6	5.04	1	0.03*	0.13
		<i>subtotal</i>	<i>14.7</i>	<i>14.1</i>	<i>0.6</i>	<i>-1.5 to 2.7</i>	<i>0.32</i>	<i>1</i>	<i>0.57</i>	
	no change	7	0.5	5.9	5.4	4.5 to 6.4	89.21	1	0.00***	2.22
		8	48.2	50.4	2.2	-0.8 to 5.1	2.13	1	0.14	
		9	10.7	7.8	2.9	1.2 to 4.7	11.36	1	0.00***	0.10
		<i>subtotal</i>	<i>59.4</i>	<i>64.2</i>	<i>4.8</i>	<i>2.2 to 8.0</i>	<i>11.41</i>	<i>1</i>	<i>0.00***</i>	<i>0.48</i>
unclear			-	0.3	-	-	-	-	-	
<i>total</i>			<i>100</i>	<i>100</i>						

When it comes to answer changes that were *not* beneficial, the two task types performed similarly, with no statistically significant difference overall. 14.7 percent of all answers were changed to no benefit in the MC tasks compared to 14.1 percent in the NF tasks. However, there were differences for two of the three categories within this type of answer change. While category 4 (no answer after first, incorrect after second) performed similarly between the two task types, with 11.3 percent for MC and 10.3 percent for NF, category 5 (incorrect after first, changed after second but still incorrect) was observed significantly more often for NF tasks (2.6 percent) than MC tasks (1.5 percent) with a large effect size (Cohen's $h = 0.91$), presumably because students have more opportunities to change open answers than MC answers. Similarly, there was a

significant difference between the two task types for category 6 (correct after first, incorrect after second), in that students would erroneously change their correct answers significantly more often in MC tasks than in NF tasks during the second play (1.9 percent compared to 1.1 percent), however, the effect size was small (Cohen's $h = 0.13$). This may again have to do with the nature of the task type: When students are not sure about their answer it is very easy for them to choose a different answer to a MC question but it takes more effort to change an open answer on a NF question.

Interesting findings also emerged for the three categories related to no changes at the bottom of Table 18. For category 7 (no answer in either first or second) a significant difference between the two task types with a large effect size (Cohen's $h = 2.22$) was observed: Only 0.5 percent of all MC answers were left blank after the second listening, compared to 5.9 percent of NF answers. This seems to be evidence that MC tasks may be more prone to guessing than NF tasks. There was no significant difference between the two task types for category 8 (correct after first, no changes after second), likely because students who are sure of their answer during the first listening do not change it during the second, regardless of test format. In contrast, category 9 (incorrect after first, no changes after second) was observed significantly more often for MC tasks (10.7 percent of all answers) than NF tasks (7.8 percent of all answers), however, the effect size was again small (Cohen's $h = 0.10$). This may confirm the task type effect observed on category 5 outlined above, i.e. that test takers have more opportunities to change open answers during the second play compared to MC answers. Overall, as outlined in the last section, students changed their answers significantly less often for NF tasks than MC tasks, with a medium effect size.

As a last step, the two NF specific answer change categories were looked at in detail (categories 2 and 3 in Table 16). These answer changes were regarded as beneficial, as they are evidence of students understanding more of the listening text during the second play, or at least an indication that students have more opportunities to show their understanding in the second play as compared to the first. The category frequencies are shown in Table 19.

For 6.4 percent of all NF answers, students added more details during the second play, whereas for 2.5 percent they chose a different correct answer (for questions which allowed more than one answer).

Table 19: Study 1: frequencies for the two NF specific answer change categories

	<i>N</i>	%
2 correct after first, more details after second and "more correct"	175	6.4
3 correct after first, different correct answer after second	69	2.5
total	244	8.9

In summary, the analysis of answer changes in the second play of double play revealed a number of benefits for test takers. Candidates changed their answers in about 40 percent of all cases for the MC tasks and 35 percent of all cases for the NF tasks. Out of these changes, about 60 percent resulted in benefits for the test takers, in that they were able to change their missing or incorrect answer to a correct answer. In the case of NF tasks, candidates were also able to add more details or choose a different correct answer in a number of cases. This indicates that for about 20 to 25 percent of all questions participants understood more of the listening text during the second play, or at least that they had more opportunities to showcase their understanding.

4.4. Questionnaire 1: strategies and anxiety

Questionnaire 1 targeted test-taking strategies and listening strategies as well as test-taking anxiety and listening anxiety. Participants had to indicate their level of agreement to 25 statements (24 for the NF tasks) on a four-point Likert scale. They completed the questionnaire twice – once after the single play condition and once after the double play condition (see Section 3.5.3).

The questionnaire data was analysed in three separate stages. First, descriptive statistics were calculated for each question. Then, an exploratory factor analysis was performed to group questionnaire responses. In a final step, differences in test takers' strategic behaviour and anxiety levels between the two conditions were explored by means of Wilcoxon signed-rank test.

4.4.1. Descriptive statistics

The questionnaire included both positively as well as negatively formulated questions. Thus, before calculating descriptive statistics for the questionnaire responses, the data for the negatively formulated questions (questions 8 to 13 and 20 to 23) was reversed to allow for cross-comparisons (see also Dörnyei & Taguchi, 2009, p. 90). After this

recoding, the mean and standard deviation were calculated and are displayed in Table 20 for the MC tasks and in Table 21 for the NF tasks (single play and double play).

Table 20: Study 1: descriptive statistics for the responses to Questionnaire 1: MC tasks in single and double play

	Single play MC			Double play MC		
	N*	M**	SD	N*	M**	SD
1. I read the questions/answer options before listening.	152	3.45	0.796	152	3.43	0.850
2. I tried to find my own answer during listening and only looked at the options at the end. [only for MC tasks]	150	1.87	0.849	147	1.78	0.840
3. I made a guess based on vocabulary used in the questions (and options).	150	1.82	0.949	147	1.69	0.873
4. I listened for the words that appeared in the questions (and options).	150	2.05	0.877	151	1.98	0.836
5. I only listened for relevant information to answer the questions.	150	2.67	0.902	146	2.57	0.953
6. I filled in the answer sheet anyway, though I wasn't not sure.	151	3.77	0.647	152	3.49	0.935
7. Before taking the test, I felt confident and relaxed.	145	3.09	0.897	145	3.08	0.898
8. During the test, I found myself thinking of the consequences of failing.	148	3.20	1.043	149	3.28	0.972
9. During the test, I got so nervous that I forgot facts I really know.	150	3.58	0.726	148	3.59	0.669
10. After taking the test, I felt I could have done better than I actually did.	144	2.62	0.967	145	2.80	0.997
11. When I first got my copy of the test, it took me a while to calm down to the point where I could begin to think straight.	148	3.03	0.979	151	3.25	0.952
12. While I took the test, my nervousness caused me to make careless errors.	141	3.42	0.855	145	3.50	0.765
13. While taking the test, I found myself wondering whether the other students were doing better than I was.	147	3.03	1.060	151	3.11	1.043
14. I concentrated hard on what the speaker was saying.	150	3.54	0.672	149	3.51	0.694
15. I guessed the meaning of unknown words, using tone of voice as a clue.	145	2.11	0.980	139	2.07	0.968
16. While listening, I made up a story line, or adopted a clever perspective.	143	2.33	0.933	148	2.47	1.039
17. I made a mental or written summary of language and information presented in the listening tasks.	149	1.73	0.905	147	1.80	0.929
18. I translated what I heard into my mother tongue.	147	1.39	0.688	149	1.84	0.966
19. While listening, I monitored my understanding of the listening passage discourse structure.	144	1.53	0.719	140	1.46	0.762
20. I got upset when I was not sure whether I understood what I was hearing in English.	151	2.79	1.043	152	2.89	1.004
21. I often understood the words but still couldn't quite understand what the speaker was saying.	149	3.31	0.861	150	3.18	0.852
22. I got so confused I couldn't remember what I'd heard.	149	3.32	0.923	145	3.43	0.873
23. I felt intimidated while listening to the tasks.	150	3.43	0.839	149	3.53	0.793
24. I enjoyed listening to the tasks.	137	1.97	0.923	141	2.02	0.890
25. I felt confident while listening to the tasks.	144	2.49	0.953	142	2.44	0.941

* This column displays the number of valid responses. The total number of respondents was 153.

** The mean is based on a four-point Likert scale where 1=disagree, 2=partly disagree, 3=partly agree, and 4=agree

Table 21: Study 1: descriptive statistics for the responses to Questionnaire 1: NF tasks in single and double play

	Single play NF			Double play NF		
	N*	M**	SD	N*	M**	SD
1. I read the questions/answer options before listening.	150	1.78	0.911	153	3.73	0.620
3. I made a guess based on vocabulary used in the questions (and options).	147	2.05	0.902	153	1.73	0.866
4. I listened for the words that appeared in the questions (and options).	144	2.75	0.957	151	2.07	0.910
5. I only listened for relevant information to answer the questions.	152	1.98	1.226	151	2.58	0.975
6. I filled in the answer sheet anyway, though I wasn't not sure.	142	2.80	1.012	152	3.00	1.190
7. Before taking the test, I felt confident and relaxed.	146	3.20	1.061	150	3.07	0.864
8. During the test, I found myself thinking of the consequences of failing.	149	3.44	0.766	147	3.29	1.020
9. During the test, I got so nervous that I forgot facts I really know.	143	2.29	1.072	150	3.55	0.681
10. After taking the test, I felt I could have done better than I actually did.	150	3.10	1.054	140	2.79	0.958
11. When I first got my copy of the test, it took me a while to calm down to the point where I could begin to think straight.	142	3.25	0.934	151	3.18	0.924
12. While I took the test, my nervousness caused me to make careless errors.	151	2.93	1.112	143	3.30	0.831
13. While taking the test, I found myself wondering whether the other students were doing better than I was.	151	3.51	0.672	147	3.07	1.025
14. I concentrated hard on what the speaker was saying.	143	2.08	0.957	152	3.66	0.563
15. I guessed the meaning of unknown words, using tone of voice as a clue.	146	2.33	1.058	144	2.37	1.069
16. While listening, I made up a story line, or adopted a clever perspective.	144	1.65	0.864	148	2.60	0.967
17. I made a mental or written summary of language and information presented in the listening tasks.	147	1.69	0.926	149	1.75	0.813
18. I translated what I heard into my mother tongue.	144	1.43	0.735	144	1.41	0.673
19. While listening, I monitored my understanding of the listening passage discourse structure.	148	2.68	1.018	140	1.64	0.814
20. I got upset when I was not sure whether I understood what I was hearing in English.	148	2.92	0.965	150	2.87	0.992
21. I often understood the words but still couldn't quite understand what the speaker was saying.	147	3.03	1.030	150	3.39	0.842
22. I got so confused I couldn't remember what I'd heard.	147	3.27	0.946	150	3.51	0.775
23. I felt intimidated while listening to the tasks.	145	1.76	0.802	147	3.57	0.712
24. I enjoyed listening to the tasks.	146	2.14	0.968	138	2.07	0.991
25. I felt confident while listening to the tasks.	150	1.78	0.911	144	2.64	1.008

* This column displays the number of valid responses. The total number of respondents was 153.

** The mean is based on a four-point Likert scale where 1=disagree, 2=partly disagree, 3=partly agree, and 4=agree.

Several tendencies can be identified in these results. First, the differences in means between the single play and double play condition across both task types are relatively small for most questions. However, for the majority of questions the differences seem to suggest that in the single play condition, test takers in general used more test-taking strategies (questions 1 to 6) and fewer listening strategies (questions 14 to 19), particularly for MC tasks. Test takers also seemed more anxious in the single play condition, both in terms of test-taking anxiety (questions 7 to 13) as well as listening anxiety (questions 20 to 25), although there are a number of questions which point in the opposite direction. In general, however, clear patterns are difficult to make out by only inspecting the means and standard deviations.

In order to get a clearer picture of possible patterns between the two conditions a factor analysis was performed. The results of the factor analysis served as the basis for a Wilcoxon signed-rank test to inspect statistical significance, as will be described in the following.

4.4.2. Factor analysis

An exploratory factor analysis was conducted separately for the two conditions (single play and double play) to group questionnaire responses in order to validate the questionnaire categories and to simplify subsequent tests for statistical differences. It was decided to join the data for the individual tasks in each condition in order to achieve a larger sample size and higher common factor variance without cross-loadings, following recommendations by Osborne and Costello (2005). The datasets for the two separate conditions were therefore responses based on both MC and NF tasks, across all participants (N=304, with 2 missing responses). For this reason, one item had to be dropped prior to the analysis as this item was only included for the MC tasks but not the NF tasks (Item 2 in the questionnaire for MC tasks). The remaining items were the same for the two tasks.

Principal axis factoring with Varimax rotation was chosen as extraction method. De Winter and Dodou suggest using principal axis factoring for data with “a relatively simple factor pattern” (2012, p. 708), which was the case as it was hypothesised that the factors would cluster according to the four sections of the questionnaire (test-taking strategies, listening strategies, test-taking anxiety, and listening anxiety,). The analyses were run with both Varimax and Direct Oblimin rotation, which yielded essentially the

same results. Only the results based on Varimax rotation are presented here, as findings obtained by Varimax rotation are easier to interpret than findings based on Direct Oblimin rotation (Osborne & Costello, 2005, p. 3).

As suggested by Osborne and Costello (2005, p. 3), the factor analysis was run several times for both conditions to 1) identify the number of factors to be included in the final analysis by inspecting the scree plot each time and 2) detect outlier items which cross-loaded onto separate factors. Kaiser-Meyer-Olkin (KMO) test for sampling adequacy and Bartlett's test for sphericity were performed for each separate analysis and were found to be adequate in each case (KMO ranged between 0.83 and 0.87 and Bartlett's test was significant at the 0.00 level each time). For both conditions, the same three main factors were detected after inspection of the scree plots and five items were identified and removed for the final analysis, as these items each cross-loaded onto different factors:

Item 1: I read the questions/answer options before listening. (test-taking strategies)

Item 6: I filled in the answer sheet anyway, though I wasn't sure. (test-taking strategies)

Item 14: I concentrated hard on what the speaker was saying. (listening strategies)

Item 18: I translated what I heard into my mother tongue. (listening strategies)

Item 24: I enjoyed listening to the tasks. (listening anxiety)

The final analysis for both datasets was run with a fixed number of three factors. KMO was 0.87 for the single play condition and 0.86 for the double play condition and Bartlett's test was significant at the 0.00 level for both conditions (Table 22). The total variance explained was 48.64 percent for the single play and 45.06 percent for the double play condition. For both conditions, the same items loaded mainly onto the same factors (Table 23), except for item 3 and item 21, which loaded mainly onto a different factor in the double play condition. However, it was decided to keep these items in the analysis as in the single play condition they loaded mainly onto the same factor. As hypothesised, the identified factors correspond to the pre-specified categories of the questionnaire. One factor relates to test-taking strategies (items 3 to 5), one to listening strategies (items 15 to 19), and one to anxiety in general (items 7 to 13 and 20 to 25);

items related to test-taking anxiety loaded onto the same factor as items related to listening anxiety.

Following this, as suggested by Dörnyei and Taguchi (2009, pp. 93–95), Cronbach’s Alpha was calculated based on the complete dataset but separately for the three identified factors to inspect reliability. For test-taking strategies Cronbach’s Alpha was .70, for listening strategies it was .75, and for anxiety it was .93, indicating that the individual factors were reliably measuring their respective constructs (see also Vogt, 2007). In addition, none of the items contributed negatively to overall reliability, that is no item would have increased Cronbach’s Alpha if deleted.

Table 22: Study 1: KMO measure of sampling adequacy and Bartlett’s test of sphericity for the factor analysis of the responses to Questionnaire 1

		Single play	Double play
KMO measure of sampling adequacy		0.87	0.86
Bartlett’s test of sphericity	approx. Chi-Square	1255.00	1062.38
	df	171	171
	<i>p</i>	0.00	0.00

Table 23: Study 1: rotated factor matrix for the responses to Questionnaire 1

	Single play			Double play			Factor
	1	2	3	1	2	3	
3. I made a guess based on vocabulary used in the questions (and options).	-0.08	0.22	0.32	-0.15	-0.03	0.18	test-taking strategies
4. I listened for the words that appeared in the questions (and options).	-0.10	-0.02	0.63	-0.21	-0.59	0.12	
5. I only listened for relevant information to answer the questions.	-0.12	-0.08	0.56	-0.15	-0.57	0.03	
7. Before taking the test, I felt confident and relaxed.	0.63	0.03	-0.14	0.64	0.15	0.13	anxiety
8. During the test, I found myself thinking of the consequences of failing.	0.62	-0.01	-0.03	0.53	0.21	-0.14	
9. During the test, I got so nervous that I forgot facts I really know.	0.70	-0.06	-0.14	0.68	0.09	-0.04	
10. After taking the test, I felt I could have done better than I actually did.	0.51	-0.12	-0.20	0.35	0.26	-0.07	
11. When I first got my copy of the test, it took me a while to calm down to the point where I could begin to think straight.	0.61	-0.15	-0.20	0.63	0.06	-0.03	
12. While I took the test, my nervousness caused me to make careless errors.	0.71	-0.11	-0.20	0.68	0.08	-0.09	
13. While taking the test, I found myself wondering whether the other students were doing better than I was.	0.50	0.07	0.13	0.46	0.27	0.03	
15. I guessed the meaning of unknown words, using tone of voice as a clue.	-0.10	0.30	0.05	-0.06	-0.05	0.40	listening strategies
16. While listening, I made up a story line, or adopted a clever perspective.	0.01	0.52	-0.20	0.10	0.04	0.41	
17. I made a mental or written summary of language and information presented in the listening tasks.	0.12	0.74	0.04	0.14	0.03	0.61	
19. While listening, I monitored my understanding of the listening passage discourse structure.	0.01	0.40	0.04	-0.03	-0.05	0.48	
20. I got upset when I was not sure whether I understood what I was hearing in English.	0.72	-0.10	-0.01	0.63	0.24	-0.02	anxiety
21. I often understood the words but still couldn't quite understand what the speaker was saying.	0.57	0.20	-0.17	0.40	0.43	0.11	
22. I got so confused I couldn't remember what I'd heard.	0.67	0.03	-0.17	0.60	0.25	0.19	
23. I felt intimidated while listening to the tasks.	0.76	-0.08	-0.04	0.67	0.20	0.03	
25. I felt confident while listening to the tasks.	0.70	0.16	-0.05	0.54	0.33	0.23	

Extraction method: principal axis factoring

Rotation method: Varimax with Kaiser normalization

Rotation converged in four iterations for the single play and five iterations for the double play condition

4.4.3. Wilcoxon signed-rank test

A Wilcoxon signed-rank test was performed based on the means of the three identified factors (test-taking strategies, listening strategies, and anxiety) to inspect differences between the single play and double play condition. As shown in Table 24, test takers relied more on test-taking strategies and less on listening strategies and they were more anxious in single play compared to double play. The differences were statistically

significant at the 0.05 level for test-taking strategies and at the 0.01 level for listening strategies and anxiety, with small to medium effect sizes (Table 25).

Table 24: Study 1: descriptive statistics for the three factors of the responses to Questionnaire 1

Pairs	N*	M**	SD
test-taking strategies – single play	304	2.18	.65
test-taking strategies – double play	305	2.10	.64
listening strategies – single play	303	1.91	.59
listening strategies – double play	304	2.02	.61
anxiety – single play	304	3.01	.67
anxiety – double play	305	3.17	.61

* This column displays the number of valid responses. The total number of respondents was 306.

** The mean is based on a four-point Likert scale where 1=disagree, 2=partly disagree, 3=partly agree, and 4=agree.

Table 25: Study 1: Wilcoxon signed-rank test and effect sizes for the three factors of the responses to Questionnaire 1

Pairs (single play and double play)	Z	Asymp. sig. (2-tailed)	Effect size <i>r</i>
test-taking strategies	-1.93	.05	0.07
listening strategies	-4.09	.00	-0.17
anxiety	-6.12	.00	-0.25

4.5. Questionnaire 2: biodata and task perception

In Questionnaire 2 students were asked about their familiarity with the topics and task types, their perceived task difficulty, and how well they were able to show their listening proficiency in each of the four tasks. The final question addressed whether participants preferred single play or double play and for what reasons. As suggested by Wagner (2012, p. 3), the median and the mode will be reported for each question because the distance between the Likert scale points for the individual questions was not regular. In addition, Mann-Whitney U tests were performed to explore statistically significant differences between the two conditions and the two sub-groups. Effect sizes were calculated following the guidelines by J. Cohen (1988). A full set of frequencies for each question is included in Appendix 10.

4.5.1. Topic familiarity

Students were asked to indicate on a 4-point Likert scale how familiar they were with the topics of the tasks (1=not at all familiar, 2=rather not familiar, 3=rather familiar,

4=very familiar). As shown in Table 26, students were less familiar with the topic of MC1 (an interview with TV director Michael Apted) than MC2 (an interview about recycling plastic bottles into school uniforms), with no significant difference between the two sub-groups. For the NF tasks, a Mann-Whitney U test revealed that students from sub-group 1 were somewhat less familiar with the topic of NF2 than the students from sub-group 2, with a small effect size ($U = 9,562.5$, $p = 0.00$, $r = 0.16$). Despite this small difference, the medians and modes for both groups show that students were generally not familiar with the topics of either NF task.

Table 26: Study 1: topic familiarity across two sub-groups and the four tasks

Sub-group	Statistics	Topic familiarity			
		MC1	MC2	NF1	NF2
sub-group 1 (N=152-153)	Median	1	3	1	1
	Mode	1	3	1	1
sub-group 2 (N=153)	Median	1	3	1	2
	Mode	1	3	1	1

4.5.2. Task type familiarity

As with topic familiarity, students were also asked to indicate on a 4-point Likert scale how familiar they were with the task types used in the test (1=not at all familiar, 2=rather not familiar, 3=rather familiar, 4=very familiar). However, the results should be interpreted with caution, as the data is clouded by the fact that it was collected after the experiment, so the influence of experiencing the task in single play or double play cannot be untangled from the students' ratings. Students may have considered the single or double play condition to be part of the task type in their assessment of task type familiarity, as they experienced the same task type in the same condition: either both MC tasks in double play and both NF tasks in single play (sub-group 1), or vice versa (sub-group 2).

The results of the analysis are displayed in Table 27. As can be seen, the students were generally very familiar with the two task types used in the study. For the MC tasks, the majority of students from both sub-groups were very familiar with the task type, however with slight differences between the two sub-groups. A Mann-Whitney U test showed that students who took the tasks in the single play condition (sub-group 2) were slightly more familiar with the MC task type than students who took the tasks in the double play condition (sub-group 1), with a small effect size ($U = 10,317$, $p = 0.02$, $r =$

0.14). A slightly more pronounced difference between the two sub-groups was observed for the NF tasks. Although the majority of students from both sub-groups again indicated to be very familiar with the task type, results from a Mann-Whitney U test indicate that students who took the NF tasks in a single play condition (sub-group 1) were slightly less familiar with the task type than students who took the tasks in a double play condition (sub-group 2), with a medium effect size ($U = 7,129.5$, $p = 0.00$, $r = 0.39$).

One possible explanation for this difference is that students might generally be less familiar with NF than MC, as MC is commonly used across all language skills and also across other subjects. In addition, students are used to a double play condition, because all listening texts in the Matura exam and hence the preparatory classroom exams during secondary education are played twice. Therefore, the students who experienced the NF tasks in single play may have felt even less familiar with this task type than the general population, as they might have taken the single play condition into account when rating for task type familiarity, resulting in the slightly lower ratings of sub-group 1. In general, however, the majority of both groups of students indicated to be “very familiar” with the NF task type.

Table 27: Study 1: task type familiarity across the two sub-groups and the two task types

Sub-group	Statistics	Task type familiarity	
		MC	NF
sub-group 1 (N=152-153)	Median	4	3
	Mode	4	4
sub-group 2 (N=153)	Median	4	4
	Mode	4	4

4.5.3. Perceived difficulty of the listening tasks

Students were also asked to indicate on a 4-point Likert scale how difficult they found the listening tasks (1=very difficult, 2=rather difficult, 3=rather not difficult, 4=not difficult). As shown in Table 28, for the MC tasks the medians and modes are all at 3, with the exception of the mode for MC1 for sub-group 2, which came out at 2. However, differences between the two groups were not statistically significant. For the NF tasks, on the other hand, Mann-Whitney U tests showed that students from sub-group 1, who experienced the NF tasks in single play, perceived the listening tasks to be significantly more difficult than students from sub-group 2, who completed the tasks in double play,

with a medium effect size for NF1 and a large effect size for NF2 (NF1: $U = 6,557.5$, $p = 0.00$, $r = 0.40$; NF2: $U = 5,076.5$, $p = 0.00$, $r = 0.51$). Importantly, the difference here spans across the divide of “rather difficult” and “rather not difficult”.

Table 28: Study 1: perceived difficulty of the listening tasks across the two sub-groups and the four tasks

Sub-group	Statistics	Perceived difficulty			
		MC1	MC2	NF1	NF2
sub-group 1 (N=153)	Median	3	3	2	2
	Mode	3	3	2	1
sub-group 2 (N=152-153)	Median	3	3	3	3
	Mode	2	3	3	3

To rule out the possibility that these results are caused by potential proficiency differences between the two sub-groups (i.e. sub-group 1 might have been less proficient than sub-group 2 overall and therefore may have perceived the tasks to be more difficult), Mann-Whitney U tests were calculated to explore differences in perceived task difficulty within each sub-group between conditions and across task types. For sub-group 1, the differences in perceived task difficulty between the two conditions and across the two task types were statistically significant for all possible pairs of tasks, with medium to large effect sizes, as shown in Table 29 and Table 30. For all possible pairs, participants from sub-group 1 perceived the tasks to be more difficult in the single play condition.

Table 29: Study 1: perceived difficulty of the listening tasks for sub-group 1

Condition	Statistics	Perceived task difficulty	
		NF1	NF2
single play (N=153)	Median	2	2
	Mode	2	1
		MC1	MC2
double play (N=153)	Median	3	3
	Mode	3	3

Table 30: Study 1: crosstabulation of Wilcoxon signed-rank tests' p-values and effect sizes on perceived difficulty of the listening tasks for sub-group 1

	NF1 (single play)	NF2 (single play)
MC1 (double play)	$p = 0.00$ $r = 0.42$	$p = 0.00$ $r = 0.66$
MC2 (double play)	$p = 0.00$ $r = 0.33$	$p = 0.00$ $r = 0.59$

For sub-group 2, the differences between the two conditions across task types are smaller than for sub-group 1 (Table 31). However, the differences in perceived task difficulty were still statistically significant for all possible pairs of tasks, with small to medium effect sizes (Table 32). As sub-group 1, participants from sub-group 2 perceived all four tasks to be more difficult in the single play condition.

Table 31: Study 1: perceived difficulty of the listening tasks for sub-group 2

Condition	Statistics	Perceived task difficulty	
		MC1	MC2
single play (N=152-153)	Median	3	3
	Mode	2	3
double play (N=153)	Statistics	NF1	NF2
		3	3
	Mode	3	3

Table 32: Study 1: crosstabulation of Wilcoxon signed-rank tests' p-values and effect sizes on perceived difficulty of the listening tasks for sub-group 2

	NF1 (double play)	NF2 (double play)
MC1 (single play)	$p = 0.00$ $r = 0.31$	$p = 0.03$ $r = 0.18$
MC2 (single play)	$p = 0.00$ $r = 0.34$	$p = 0.03$ $r = 0.18$

These results suggest that students perceive tasks to be more difficult in single play compared to double play, which confirms the findings on average task difficulty described in Section 4.2.

4.5.4. Face validity

Another question explored how well students were able to show their listening proficiency through the tasks used in the study (face validity) on a 4-point Likert scale (1=not well at all, 2=rather not well, 3=rather well, 4=very well). As shown in Table 33, for the MC tasks the results are similar for both sub-groups, however, for MC1 a Mann-Whitney U test showed that students from sub-group 2, who experienced the MC tasks in single play, rated face validity significantly lower than students from sub-group 1, who completed the MC tasks in double play, with a small effect size ($U = 9,832.5$, $p = 0.02$, $r = 0.14$).

Table 33: Study 1: face validity across the two sub-groups and the four tasks

Sub-group	Statistics	Face validity			
		MC1 (double)	MC2 (double)	NF1 (single)	NF2 (single)
sub-group 1 (N=151)	Median	3	3	2	2
	Mode	3	3	2	2
sub-group 2 (N=153)	Statistics	MC1 (single)	MC2 (single)	NF1 (double)	NF2 (double)
		3	3	3	3
	Mode	2 and 3	3	3	3

For the NF tasks the differences between the two sub-groups were again more pronounced. Students who took the tasks in the double play condition (sub-group 2) felt that they were better able to show their listening proficiency compared to students who took the tasks in the single play condition (sub-group 1). Differences were statistically significant with medium effect sizes (NF1: $U = 5,709.5$, $p = 0.00$, $r = 0.46$; NF2: $U = 5,564.5$, $p = 0.00$, $r = 0.47$).

Similar to perceived difficulty, the analyses were also run separately for each sub-group across task types to rule out noise caused by potential differences between the two groups of students. As displayed in Table 34 and Table 35, students from sub-group 1 perceived the MC tasks in double play to be more valid than the NF tasks in single play, with medium to large effect sizes.

Table 34: Study 1: face validity for sub-group 1

Condition	Statistics	Face validity	
		NF1	NF2
single play (N=153)	Median	2	2
	Mode	2	2
		MC1	MC2
double play (N=153)	Median	3	3
	Mode	3	3

Table 35: Study 1: crosstabulation of Wilcoxon signed-rank tests' p-values and effect sizes on face validity of the listening tasks for sub-group 1

	NF1 (single play)	NF2 (single play)
MC1	$p = 0.00$	$p = 0.00$
(double play)	$r = 0.47$	$r = 0.63$
MC2	$p = 0.00$	$p = 0.00$
(double play)	$r = 0.47$	$r = 0.60$

The results for sub-group 2 show the same trend. Although differences were slightly smaller than for sub-group 1 (Table 36), students from sub-group 2 also felt that they were better able to show their listening proficiency in double play (NF tasks) than single play (MC tasks), with medium effect sizes (Table 37).

Table 36: Study 1: face validity for sub-group 2

Condition	Statistics	Face validity	
		MC1	MC2
single play (N=152-153)	Median	3	3
	Mode	2 and 3	3
		NF1	NF2
double play (N=153)	Median	3	3
	Mode	3	3

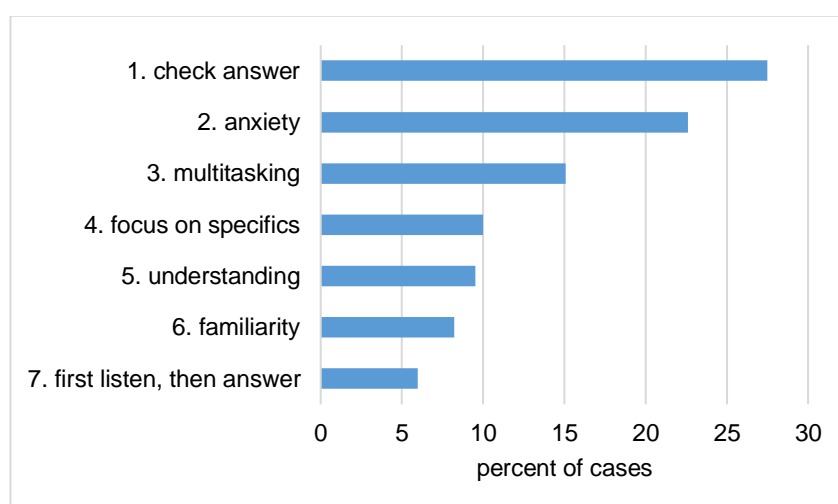
Table 37: Study 1: crosstabulation of Wilcoxon signed-rank tests' p-values and effect sizes on face validity of the listening tasks for sub-group 2

	NF1 (double play)	NF2 (double play)
MC1	$p = 0.00$	$p = 0.00$
(single play)	$r = 0.44$	$r = 0.26$
MC2	$p = 0.00$	$p = 0.01$
(single play)	$r = 0.41$	$r = 0.21$

4.5.5. Preference for single or double play

In the final question the students were asked whether they prefer single play or double play and to give reasons for their preference in an open answer. The great majority of students (98.4%) answered that they prefer double play, with one missing response. The open answers were analysed in Atlas.ti and grouped into seven categories. Figure 18 displays the frequencies of the seven coding categories in percent of cases. Each of the categories is described in detail with examples below the figure. The students' comments were in German but are translated here for ease of reading.

Figure 18: Study 1: coding categories and frequencies for the open question in Questionnaire 2 ("Why do you prefer double play?")



1. **check or correct answer or add information** (124 comments):

The students reported that they used the second play to check or correct their answers or to add extra information in the NF tasks.

Participant 516: [During the second play] I can check my answers.

Participant 416: [During the second play] I can add stuff.

2. **anxiety** (102 comments):

In the single play condition, the students felt panicked, were stressed, under pressure, or nervous, which either inhibited their listening ability or made them feel insecure or uncomfortable.

Participant 1401: In single play I'm panicked from the beginning because I might miss an answer.

Participant 802: I feel calmer and less stressed during the first play [in a double play condition].

3. **multitasking** (68 comments):

In the single play condition, the students struggled with doing multiple things at the same time (listen, read the questions, think about the answers, write down the answers, check the answers) and therefore they missed some information.

Participant 111: Because while I'm writing I can't focus on listening for the next question.

Participant 604: Because you can't focus on everything at the same time and you overhear some words [...] (overhear meaning to miss words – “überhören”).

4. **focus on specific questions, passages, or words** (45 comments):

The students used the second play to focus particularly on unanswered questions or specific passages or words.

Participant 617: Because then I don't need to focus on the details during the first play and I can answer the difficult questions during the second play.

Participant 1005: During the second listening I can focus more easily on the small details.

5. **general understanding** (43 comments):

The students did not understand everything during the first play and therefore needed the second play.

Participant 170: Because I don't understand everything during the first play.

Participant 1601: Much better understanding during the second play.

6. **familiarity with the text structure, topic, accent, or speed of delivery** (37 comments):

During the second play the students were more familiar with the listening text's structure (i.e. the order of the questions and where in the recording the relevant information can be found), or they were more familiar with the topic, the accent, or the speed of delivery.

Participant 609: During the first time I familiarise myself with the text.

Participant 1105: I need the first play to get used to the topic and the accent.

7. **first play listen, second play answer** (27 comments):

The students focussed on listening during the first play and on answering the questions during the second play.

Participant 106: During the first play I listen hard and during the second play I answer.

Participant 1011: The first time I try to understand the text well.

4.6. Summary of the main findings

Overall, the results for Study 1 show that double play not only impacts the statistical performance of test items, but also test takers' response processes, and that a number of the observed effects are positive in terms of construct validity. First, the CTT results revealed that item discrimination and overall reliability are enhanced by the double play condition compared to the single play condition, regardless of test format. Second, the Facets bias analyses showed that test items get significantly easier in double play versus single play, MC tasks more so than NF tasks. Significantly higher average item estimates were also observed in single play compared to the first play of double play, and in the second play of double play compared to the first play. This indicates that double play is not simply a repetition of a single play condition, but rather that test takers are behaving differently from the beginning. It was also shown that double play increased the comparability between MC and NF, as the difference in task difficulty between the two task types was smaller than in single play. Third, a detailed analysis of answer changes in the second play of double play revealed a number of benefits for test takers. Although in about 60 percent of all answers for the MC tasks and 65 percent for the NF tasks there was no answer change of the initial answer, for the remaining answers it was shown that participants used the second play to revise their responses and, in the case of NF tasks, to add more details or choose a different correct answer. This indicates that for about 20 to 25 percent of all questions participants understood more of the listening text during the second play, or at least that they had more opportunities to showcase their understanding.

The questionnaire results are further evidence that double play might be beneficial for construct validity. For Questionnaire 1, which targeted participants' strategic behaviour and anxiety levels, a Wilcoxon signed-rank test was run based on an exploratory factor analysis of questionnaire items. It was shown that test takers used fewer listening strategies and more test-taking strategies, and that they were more anxious in single play compared to double play, with small to medium effect sizes. In addition, the results for Questionnaire 2 show that participants perceived the tasks to be less difficult and they felt that were better able to show their listening proficiency in double play compared to single play, regardless of test format. Participants also clearly preferred the double play condition. They indicated that they used the second play to check their answers, that they were less anxious in double play, that they found it

difficult to cope with multiple modalities in single play (simultaneous reading of the questions, writing the answers, and listening to the text), and that they became more familiar with the listening text and the accents in the second play of double play.

5. Results Study 2

The results of Study 2 are presented in six main sections. In the first four sections, the findings on the four main response processes of interest are outlined in turn, with cognitive processes in Section 5.1, listening strategies in Section 5.2, test-taking strategies in Section 5.3, and anxiety in Section 5.4. Within each of these sections, the individual categories of the response process are first exemplified with quotations from the verbal recall data, before the results of the data analysis are outlined. The chapter then goes on to describing and analysing the additional categories which emerged during the coding in Section 5.5. Finally, Section 5.6 presents a summary of the main findings.

5.1. Cognitive processes

Evidence for four of the five cognitive processes was found in the data: the lower-level processes lexical search and parsing and the higher-level processes meaning construction and discourse construction. In the next sections, examples from the data for each of the identified cognitive processes will be presented first, followed by an analysis of the how frequently the processes were used across the different conditions (single play and double play, RQ 2), the two task types, and the four tasks (RQ 2a).

5.1.1. Examples from the data

In this section exemplary quotations related to the four identified cognitive processes lexical search, parsing, meaning construction, and discourse construction are presented. In total, 366 quotations were coded as cognitive processing across all participants, tasks, conditions, and stages of recall. In terms of the nature of processing, no discernible differences could be identified between single and double play for any of the identified cognitive processes.

Out of the 366 quotations, 150 were coded as **lexical search**. Lexical search is characterised by the recognition of individual words and is informed by lexical knowledge. This process was observed for all participants (a summary table displaying the frequencies of observed response processes for each of the 16 participants is included in Appendix 11). The following are typical examples from the data. In all instances of lexical search, candidates were listening for specific words:

P10 NF2 twice while-listening second⁴

And then I somehow/then he said “highlander”.

P02 NF1 once while-listening

So here I first heard “fishing” [...]

Like lexical search, **parsing** is also considered a lower-level listening process, as it is not associated with building meaning in context. It occurs when the listener puts individual words into a syntactic pattern and is informed by syntactic knowledge. The output of parsing is the bare meaning of an utterance at clause or sentence level. Parsing was assigned to a total of 116 quotations and was displayed by all 16 participants, for example:

P14 NF2 once while-listening

And then he said “connect to the bricks” [...]

P15 MC2 twice while-listening first

Here I understood “the blazer is”, and then [...]

In contrast to lexical search and parsing, which operate at lower-level, **meaning construction** is considered a higher-level cognitive process. Listeners conceptualise the literal meaning of utterances and relate them to the context in which they occurred to construct higher-level meaning. Meaning construction is informed by pragmatic knowledge and external knowledge about the world, the speaker, and the topic. It was observed for 15 of the 16 participants and applied to 91 quotations overall, similar to the following:

⁴ The excerpts are labelled according to the data segments described in Table 9 on page 80 and include the participant number (P01-P16), the task the excerpt is based on (MC1, MC2, NF1, NF2), the number of times the participant heard the excerpt (once, twice), and the stage of recall (retrospective, pre-listening, while-listening, post-listening, post-hoc). Thus, this particular excerpt stems from participant 10 (“P10”), who completed note form task 2 (“NF2”) in double play (“twice”), and the data was taken from their stimulated recall during the while-listening stage of the second play (“task second”).

P09 MC2 twice while-listening first

There he asks about the price. And then he says that it's similar to the normal blazer [...] But then they ask her whether it is comfortable and she says that it is very comfortable.

P13 MC1 twice while-listening second

Yes, so she was a difficult girl when she was 7 [...] but they did not say that she had mental problems, but it was clear / he also said that she improved.

The second higher-level process is **discourse construction**. It is characterised by relating the meaning of the message to the text as a whole and is informed by external knowledge about the text type, the world, and the speaker. Discourse construction was displayed by 7 participants and assigned to a total of 9 quotations, such as the following:

P01 MC1 once retrospective

It was about a man who made a TV show and he followed children of a certain age to another age and he documented their lives. And he asked them "What are your dreams?" and so on. And then he talked about how the people changed [...] and that it was extremely fascinating for him how the characters changed over time.

5.1.2. Analysis of cognitive processes in single and double play

In order to analyse to what extent the participants engaged in different levels of cognitive processing between single play and double play (RQ 2), the stimulated recall data was first inspected separately for the different stages of recall to be able to objectively compare the two conditions. The vast majority of quotations related to cognitive processing (341 out of 366) stemmed from the while-listening period: the time during which the participants listened to the texts and engaged with the tasks. The remaining 25 quotations emerged during the retrospective recalls (16 quotations) and the post-listening stage (the time students were given to check their answers after the listening text finished: 9 quotations). Thus, the main part of the analysis for cognitive processing draws on data from the while-listening period.

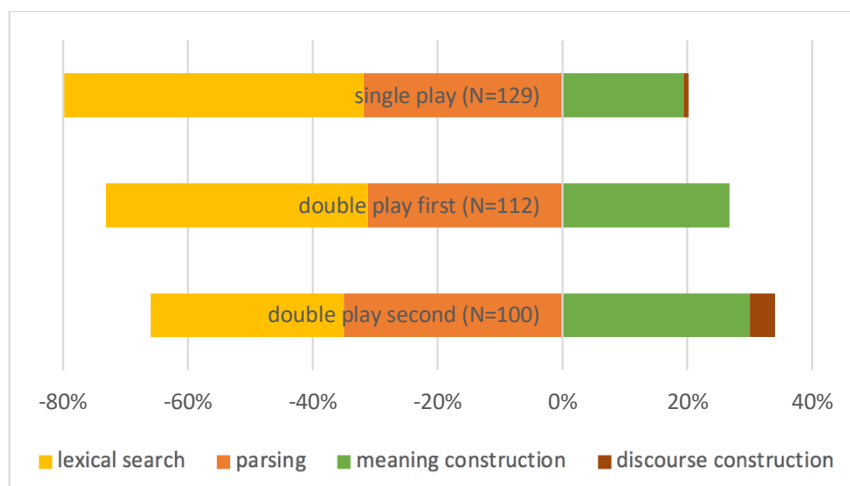
The analysis of the data was run in three steps. First, the stimulated recalls were analysed jointly for all tasks to get an overall picture of the differences in cognitive processing between single play and double play (RQ 2). This analysis was performed twice: once for the while-listening stage and once for the retrospective recalls and post-listening stage. Second, the data from the while-listening stage was analysed separately for each task type (MC and NF) to discern whether cognitive processing levels were affected by task type (RQ 2a). Finally, the four tasks were analysed individually to detect whether task type effects are in line with individual task effects. This analysis was again only based the while-listening stage. An analysis based on individual task types or tasks was not performed for the retrospective recalls and post-listening stage due to the small sample size for quotations associated with these stages of recall.

Figure 19 displays the results for the while-listening period for all tasks jointly. The chart shows the amount of cognitive processing in percent of cases for each level (lexical search, parsing, meaning construction, discourse construction). To illustrate the extent to which the proportions of lower level and higher level processing shifted between the different conditions, lower level processes (lexical search and parsing) were placed in the negative range on the x-axis, while higher level processes (meaning construction and discourse construction) were placed in the positive range on the x-axis. Each bar represents 100% of cognitive processing for the respective condition. The bar on top of the figure (single play) relates to the single play condition and the remaining two bars to the double play condition. The double play condition was split into first play (double play first) and second play (double play second) to study possible differences in behaviour between the two plays. The raw frequencies (i.e. the total number of quotations) are included in brackets for each bar.

As shown in Figure 19, in both conditions the majority of observed cognitive processes were related to lexical search and parsing. However, there was a clear difference between the three plays: In single play, 80% of observed cognitive processes were associated with lower-level processing (48% lexical search and 32% parsing) and 20% with higher-level processing (19% meaning construction and 1% discourse construction), whereas in the first play of double play, the proportions shift to 73% lower-level processing (42% lexical search and 31% parsing) and 27% higher-level processing (all meaning construction). This trend is amplified in the second play of the double play condition, where only 66% of observed processes were lower-level (31% lexical search and 35% parsing) but 34% were higher-level (30% meaning construction

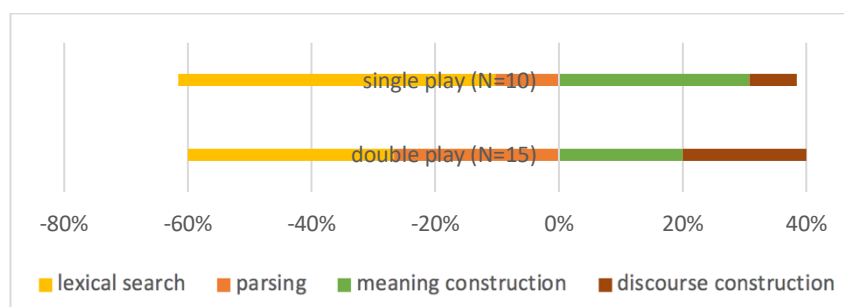
and 4% discourse construction). Thus, overall, the results suggest that test takers engaged more in meaning-building processes in double play as compared to single play, particularly during the second play of the double play condition.

Figure 19: Study 2: cognitive processes while listening for single play and double play (all tasks)



Despite the small sample size, the data from the retrospective recalls and the post-listening stage shows the same trend, as displayed in Figure 20 (the width of the bars was adjusted to reflect the small sample size for quotations compared to the while-listening period). Although the difference between the two conditions is less pronounced, overall participants reported increased levels of higher-order processing during double play as compared to single play. In the double play condition, candidates engaged in higher levels of parsing as compared to lexical search and in higher levels of discourse construction as compared to meaning construction.

Figure 20: Study 2: cognitive processes from the retrospective recalls and the post-listening stages for single play and double play (all tasks)



The following quotations illustrate the phenomenon that participants were more focussed on understanding details and specifics in single play compared to double play (participant 14), but also in the first play of double play compared to the second play (participant 4):

P14 NF2 once retrospective

So now [in single play] I was listening in more detail. [...] So I looked at the answers and listened and I tried to, like, hear the exact number of hours he needed. And I tried to pay attention to specifics, like, how much this is. [I was] listening more to the details than before.

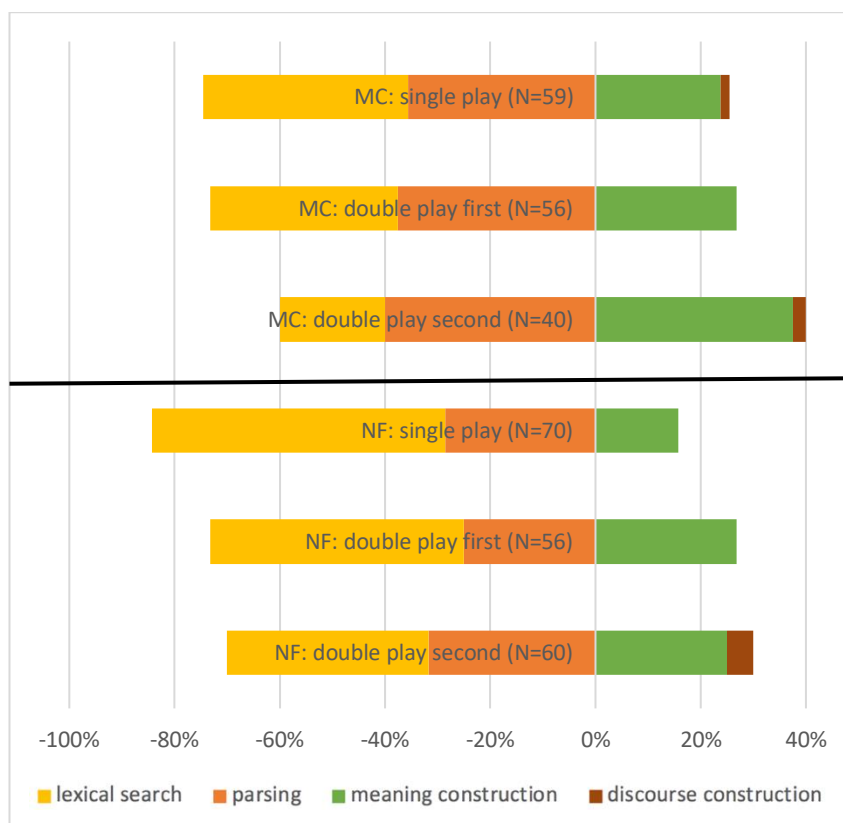
P04 NF1 twice while-listening second

[...] so it's mostly, like, [...] during the first play I listen more for the specific things, so the things I'm waiting for all the time and during the second play I understand more of the context and the other details [...] which are less important for the questions, but more important for the overall meaning.

In order to answer RQ 2a (differences in cognitive processing between the two task types), the data for the while-listening period was analysed separately for the two task types (Figure 21). As shown in the figure, the same trend of increased levels of higher-order processing in double play can be observed, with slight differences between the two task types. For the NF tasks, 16% of all observed processes in single play were higher-level (all meaning construction), compared to 27% in the first play of double play (again all meaning construction) and 30% in the second play (25% meaning construction and 5% discourse construction). For the MC tasks, on the other hand, the shift towards higher-level processing was only observed during the second play in the double play condition. For MC tasks in single play, 26% of all observed processes were higher-level (24% meaning construction and 2% discourse construction), compared to 27% during the first play in double play (all meaning construction) but 40% during the second play (38% meaning construction and 3% discourse construction). This suggests that in terms of cognitive processing for MC tasks, test takers behaved similarly in the single play condition and the first play of the double play condition, however, they engaged in increased levels of higher-order processing during the second play of the double play condition. It can also be seen that across all three plays test takers in general

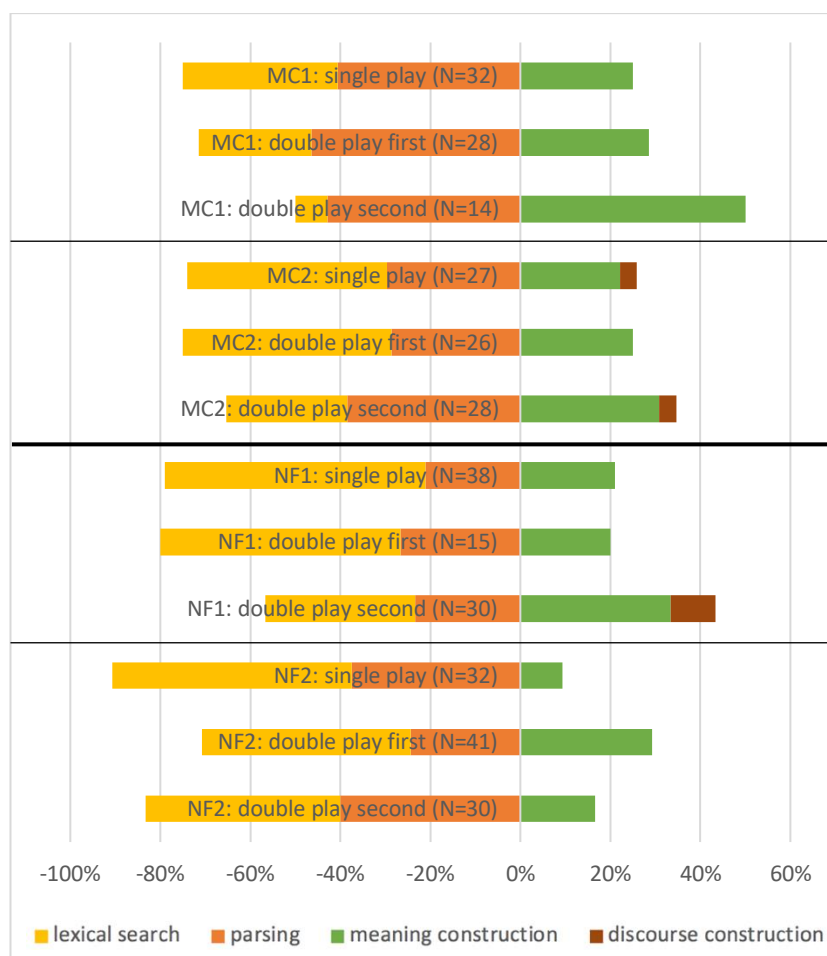
displayed more higher-level processing for MC tasks (M=30.74%) compared to NF tasks (M=24.17%), which was to be expected as the MC items targeted main ideas and supporting details whereas the NF tasks targeted specific information and important details (see Section 3.4.2.2).

Figure 21: Study 2: cognitive processes while listening for single play and double play (MC and NF)



As a last step, the four tasks were analysed individually to discern whether the detected task type effects are consistent with the findings on task level. This analysis is again based on the data from the while-listening period. As shown in Figure 22, the shift towards higher-order processing during the second play of the double play condition described for MC tasks above was evident for three of the four tasks used in the study: MC1, MC2, and NF1 all show this phenomenon, with MC1 displaying the strongest effect (a 25% increase between single play and the second play of double play), followed by NF1 (a 21% increase) and MC2 (a 10% increase). For these three tasks, test takers were generally listening more locally during single play and the first play of double play but more globally during the second play of double play.

Figure 22: Study 2: cognitive processes while listening for single play and double play (MC1, MC2, NF1, NF2)



The following excerpt of participant 4, who completed NF1 in double play, is a typical example of this phenomenon of different listening behaviour across condition:

P04 post-hoc

So, [during the first play] I don't have time to think whether my answers make sense, that's why I like the second play. Also, [during the first play] I am so focused on answering the questions, so [during the second play] I understand things which I don't understand when I answer. [...] When I'm not focusing on answering questions I understand more of the context.

Another example is the following quotation from the stimulated recall of participant 5 during the second while-listening stage of MC1 in double play. This participant had to cough during the first play and thus missed some information,

however, they also indicate that they were more focussed on understanding the meaning of the listening text during the second play compared to the first:

P05 MC1 twice while-listening second

So now [during the second play] I was more focussed on what he was saying and during the first play I was more, like, reading the questions and stuff. I also had to cough during the first play. So now [during the second play] I was able to understand more [...]

Interestingly, however, for NF2 the test takers' approach was different. The increase in higher-order processing between single play and double play was stronger during the first play of the double play condition. During the second play test takers seemed to fall back on lower-level processing, albeit still to a lesser degree than during single play. One possible explanation for this difference is the fact that NF2 was based on a shorter listening text and was more difficult than the other three tasks (see Section 3.4.2.2), which might have impacted test takers' response processes. The following excerpt is taken from the post-hoc interview of participant 12, who completed NF2 in double play:

P12 post-hoc

INT: What do you prefer, single play or double play?

P12: Double play. Definitely.

INT: Why?

P12: Yes, because it's always, like, during the first play [...] I don't pay attention to the details, but more to the context and then I know what to look out for during the second play.

Despite these differences the general tendency was the same across all tasks: Participants engaged in increased levels of higher-order processing when they completed the tasks in a double play condition compared to a single play condition, and within the double play condition, higher-order processes were typically more prevalent on the second play versus the first.

5.2. Listening strategies

Evidence for 11 different listening strategies was found in the data: planning, focussing attention, monitoring, evaluation, inferencing, elaboration, prediction, contextualization, translation, and managing emotions. In the following, exemplary quotations for each of the strategies will be presented first, before analysing how the participants used these listening strategies in single play and double play (RQ 3). In order to inform RQ 3a, differences between the two task types and the four tasks were also explored.

5.2.1. Examples from the data

Overall, 421 quotations in the data were coded as a listening strategy. Out of these, 59 quotations were classified as **planning**. Planning is characterised by developing a plan for listening task completion and was mostly evident in the pre-listening stage. In a smaller number of instances participants formed a plan on how to complete the remainder of a listening task in the while-listening stage, particularly during the first play of the double play condition. Planning was observed for all 16 participants (see Appendix 11). The following quotations, taken from the stimulated recalls during the pre-listening stage in single play for participant 1 and the first while-listening stage in double play for participant 2, are typical examples:

P01 MC1 once pre-listening

Yes here I was thinking, great 60 seconds to check my answers, then I will be able to correct a few things [...] I can go through everything again and remember.

P02 NF2 twice while-listening first

[...] and here I thought I need to pay more attention to this when I hear it the second time.

The most common listening strategy was **focusing attention**, with 89 quotations overall. It is characterised by “[a]voiding distractions and heeding the auditory input in different ways, or keeping to a plan for listening development” (Vandergrift & Goh, 2012, p. 277). Focussing attention was displayed by all 16 participants and was often observed in combination with test-management, for example when participants were

listening for specific words from the test questions or answer options (selective attention):

P06 NF1 twice while-listening second

INT: How were you listening?

P03: I was more, like, listening until I hear the word “problem” or “trouble” [from the test questions].

In other cases participants were directing their attention to more general aspects, for example by focusing on the listening text instead of the test paper (directed attention), as this quotation from participant 5 during the second play of double play illustrates:

P05 MC1 twice while-listening second

I focused more on what he was saying, because during the first play I was more, like, reading through the questions.

Monitoring was observed for 15 participants and assigned to a total of 67 quotations. This listening strategy was used by participants for “checking, verifying, or correcting [their] comprehension or performance in the course of [the tasks]” (Vandergrift & Goh, 2012, p. 278). In most instances of monitoring students checked their performance or comprehension by referring to the test questions, so these quotations were also coded as test-management, as the following examples show:

P08 NF1 once while-listening

It took a relatively long time again, so I was thinking that I had missed this, because I looked at the sentence below [...] and then I thought that I had missed it [...] but then I realised ok they had not said anything about the teams and the boats yet [...] so I just continued to listen.

P13 MC1 twice while-listening first

Here I was thinking that he is talking about this question again, but because it says “Neill” here I realized that he was talking about him.

Evaluation was the second most common listening strategy, applied by all 16 participants with 75 quotations overall. Vandergrift and Goh define it as “[c]hecking the outcomes of listening comprehension or a listening plan against an internal or an external measure of completeness, reasonableness, and accuracy” (Vandergrift & Goh, 2012, p. 278). Similar to monitoring, evaluation was mostly evident in combination with test-management: Students evaluated their comprehension by checking the answers they gave (performance evaluation):

P11 MC1 twice post-listening second

So here I quickly [...] checked the answers.

P12 NF2 once post-listening

INT: And what were you thinking when the recording finished playing?

P12: Yes here I was again thinking about [my answer to] question 5, but I wasn't sure.

Inferencing was observed for 12 participants and assigned to a total of 26 quotations. This listening strategy was assigned whenever participants were “[u]sing information within the text or conversational context to guess the meanings of unfamiliar language items associated with a listening task, to predict content and outcomes, or to fill in missing information” (Vandergrift & Goh, 2012, p. 279):

P04 NF1 twice while-listening second

The second time I still did not quite understand it. I did understand “table”, so when I understood “table” I was thinking that they were probably eaten.

P07 MC2 twice while-listening first

Here I was not quite sure, because [...] it was never mentioned directly, but, um, what they meant was that it brought back discipline.

Overall, 25 quotations by 9 different participants were coded as **elaboration**. According to Vandergrift and Goh, elaboration is characterised by “[u]sing prior knowledge from outside the text or conversational context and relating it to knowledge gained from the text or conversation in order to embellish one's interpretation of the

text” (Vandergrift & Goh, 2012, p. 280). Typical examples from the data are the following:

P01 MC2 twice while-listening second

Here at “put into places”, I once watched a documentary about a posh boarding school where the headmaster wanted students to dress well so they don’t have such a, like, chill-feeling when they are at school, but a real performance-feeling. And that’s what I remembered. And I thought, that makes sense, I can also translate this situation to the listening text.

P10 NF2 twice while-listening second

And then [...] he said “highlander” and I, like, does he mean the TV show which is set in Scotland, or some other kind of “highlander”?

Prediction was used by all 16 participants and assigned to a total of 42 quotations. It was evident whenever participants tried to predict the contents of the listening text and was mostly observed in the pre-listening stage, when participants anticipated what the listening text would be about by looking at the title of the task or reading through the test questions and answer options. Therefore, in most cases prediction co-occurred with test-management:

P03 MC1 twice pre-listening first

So here I also had a look at what it was going to be about. I, like, read through everything and then had a look at what this was going to be about.

P06 NF2 once pre-listening

Yes so I looked at “Lego”, so [...] I knew this was going to be something about “building blocks” or something like that.

The listening strategy **contextualization** was only used by 6 participants and applied to a total of 8 quotations. It is characterised by “[p]lacing what is heard in a specific context in order to prepare for listening or assist comprehension” (Vandergrift & Goh, 2012, p. 282):

P01 MC2 twice while-listening second

Ok, I think I did not understand the word “principals”. But when she started talking about school and also about school uniforms and such, that made it clear for me that it must have to do with school and not with some other workplace. So somehow I understood the context but I did not understand “principals”.

P08 NF2 twice while-listening second

Here I did not want to listen for the answer, but rather see whether what I understand fits in this context.

The least common listening strategy across all participants was **translation**, with a total of 5 quotations across 4 participants. It was assigned to passages where participants translated what they heard into their first language (German) to assist comprehension:

P01 MC1 once retrospective

INT: Ok. And how were you listening?

P01: [...] I was, like, listening and then I tried to process it, to translate it or to understand it.

P08 NF2 twice while-listening second

So here they said “eight to ten hours” [and not “ten to eight hours”] and you also say it that way in German, so [...]

Finally, the listening strategy **managing emotions** was displayed by 9 participants and assigned to 25 quotations overall. Vandergrift and Goh define it as “[k]eeping track of one’s feelings and not allowing negative ones to influence attitudes and behaviors” (Vandergrift & Goh, 2012, p. 284). In all instances of this strategy, participants commented on having positive feelings about the listening exercise (self-encouragement):

P03 MC1 twice pre-listening

Yes so because I knew that I would hear it twice I thought that I did not need to stress myself about understanding [everything].

P06 NF1 twice pre-listening first

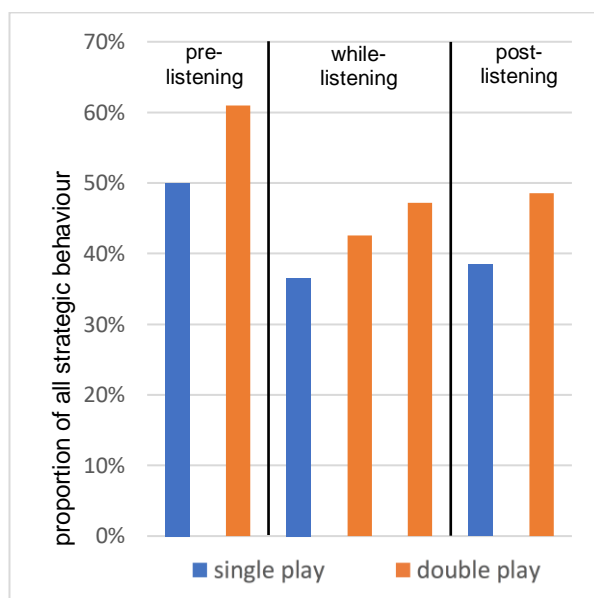
When I can hear it twice I [tell myself that I] can take my time [...] and it's not so stressful in case I don't understand something.

5.2.2. Analysis of listening strategies in single and double play

The data for listening strategies was analysed separately for the two conditions according to each stage of task completion (pre-listening, while-listening, and post-listening). To be able to objectively compare the amount of listening strategies across the different stages, the number of listening strategies was calculated as a proportion of all strategic behaviour for each stage of task completion. In other words, the number of quotations related to listening strategies was divided by the total number of quotations related to listening strategies and test-taking strategies for each stage of task completion. The analysis was first run for all tasks jointly to inform RQ 3, and then separately for the two task types and the four individual tasks to answer RQ 3a.

Figure 23 displays the total amount of listening strategies as a proportion of all strategic behaviour for each stage of task completion, for all tasks jointly. The chart juxtaposes the two conditions (single play in blue and double play in orange) for each stage: pre-listening (the time students were given to study the task before the listening text started playing), while-listening (for double play the graph displays both the first play and the second play), and post-listening (the time students were given to check their answers after the listening text finished playing). As shown in the graph, there was a clear tendency across all stages of task completion: Participants reported a relatively greater amount of listening strategies in the double play condition compared with the single play condition. In the pre-listening stage, 50% of all strategic behaviour was coded as a listening strategy in single play, compared to 61% in double play. In the while-listening stage, levels of listening strategies increased from 37% in single play to 43% in the first play of double play and to 47% in the second play of double play. Finally, in the post-listening stage, 39% of students' strategic behaviour was related to a listening strategy in single play, compared to 49% in double play.

Figure 23: Study 2: total number of listening strategies as a proportion of overall metacognitive processing for single play and double play (all tasks; by stage of task completion)



Next, the use of the individual listening strategies was analysed separately for each stage of task completion. As shown in Figure 24, in the pre-listening stage participants used three different listening strategies (planning, prediction, and managing emotions). Participants relied more on planning during single play (28% of all strategic behaviour in the pre-listening stage) than double play (20%), however, they tried to predict the topic of the listening text more often during double play (31% of all metacognitive processing in the pre-listening stage) than single play (22%). Finally, the strategy of managing emotions (exclusively self-encouragement – see discussion in Section 5.2.1 above) was only observed during the double play condition (10% of all strategic behaviour in the pre-listening stage).

Figure 24: Study 2: individual listening strategies as a proportion of overall metacognitive processing for single play and double play (all tasks; pre-listening only)

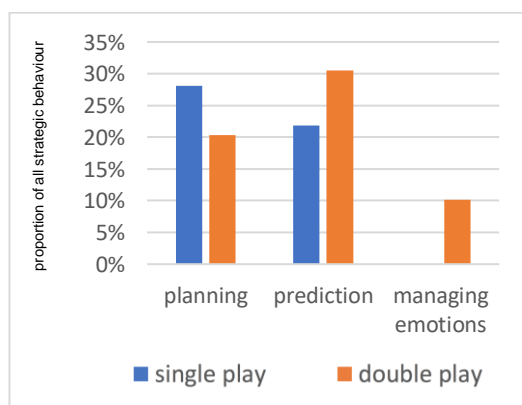
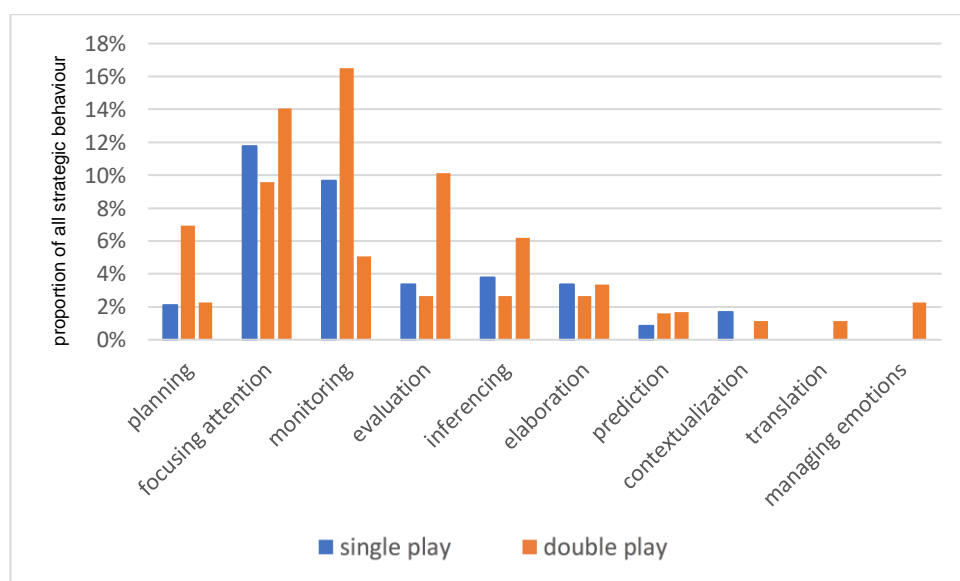


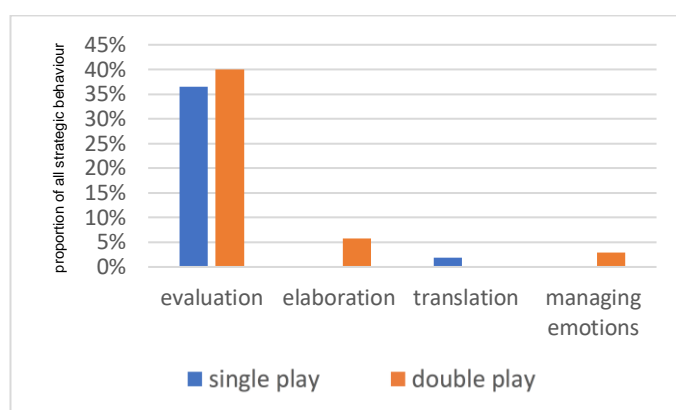
Figure 25 displays the results for the while-listening stage. In single play, participants engaged in 8 out of the 10 observed listening strategies, whereas in double play they used all 10 strategies. It can also be seen that the use of all but two listening strategies was higher in double play than single play: With the exception of contextualization and elaboration, which were observed slightly more often in single play, the amount of all listening strategies was relatively higher in double play. Within the double play condition, levels of listening-strategic behaviour were greater during the second play for the majority of strategies. Although participants relied on higher levels of planning and monitoring during the first play compared to the second play of double play, they used higher levels of focusing attention, evaluation, inferencing, elaboration, prediction, contextualization, translation, and managing emotions during the second play.

Figure 25: Study 2: individual listening strategies as a proportion of overall metacognitive processing for single play and double play (all tasks; while-listening only)



The same trend was also observed in the post-listening stage, albeit to a lesser degree, as shown in Figure 26. Evaluation was the most important listening strategy for participants after listening to the text, with 37% of all strategic behaviour in single play and 40% in double play. In double play, two students also engaged in elaboration (6%) and one student in managing emotions (3%), whereas in single play one student also used translation (2%).

Figure 26: Study 2: individual listening strategies as a proportion of overall metacognitive processing for single play and double play (all tasks; post-listening only)



Overall, these results indicate that participants engaged more with the listening text during double play versus single play, particularly during the second play in the while-listening stage, by displaying a greater variety and a greater relative amount of listening-strategic behaviour. The following excerpt from the post-hoc interview of participant 10 illustrates this phenomenon:

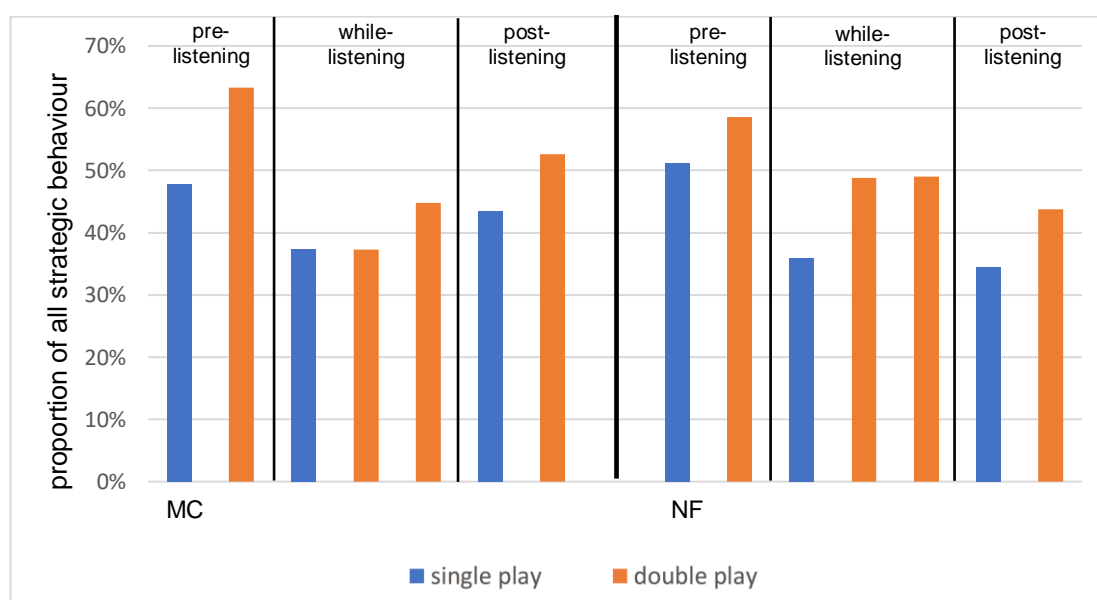
P10 post-hoc

Yes, so [when I can hear it twice] I, like, [...] think more about it. I kind of let the text get to me, [...] like, I engage more with the text. When I hear it only once I have the feeling of, like, being at war with the text and [when I hear it twice] it's much nicer, so that I have the feeling ok now I'm working with the text.

In order to investigate whether there were differences in the use of listening strategies between the two task types (RQ 3a), the data were analysed separately for MC and NF. This analysis was only run on the total number of listening strategies. A more-fine grained analysis of individual listening strategies was not performed on the level of task types as the small number of quotations for each individual listening strategy would not have allowed for meaningful comparisons. Figure 27 shows the total amount of listening strategies as a proportion of all strategic behaviour for each stage of task completion, separately for MC and NF. As can be seen in the figure, the two task types behaved similarly in terms of listening-strategic behaviour displayed by the participants. For both task types, the overall trend of increased levels of listening strategies in double play versus single play was found across all stages of task

completion. However, a difference was observed for the while-listening stage: For NF tasks, students used listening strategies in 36% of all metacognitive processing in single play, compared to 49% during both the first and second play of double play. For MC tasks, on the other hand, levels of listening-strategic behaviour were similar between single play and the first play of double play (37% of all processing for both stages), but markedly higher during the second play of double play (45%). This finding parallels the results on cognitive processes for the while-listening period of MC tasks presented above, where increased levels of higher-order cognitive processing were only found for the second play of double play compared to single play, but not for the first play.

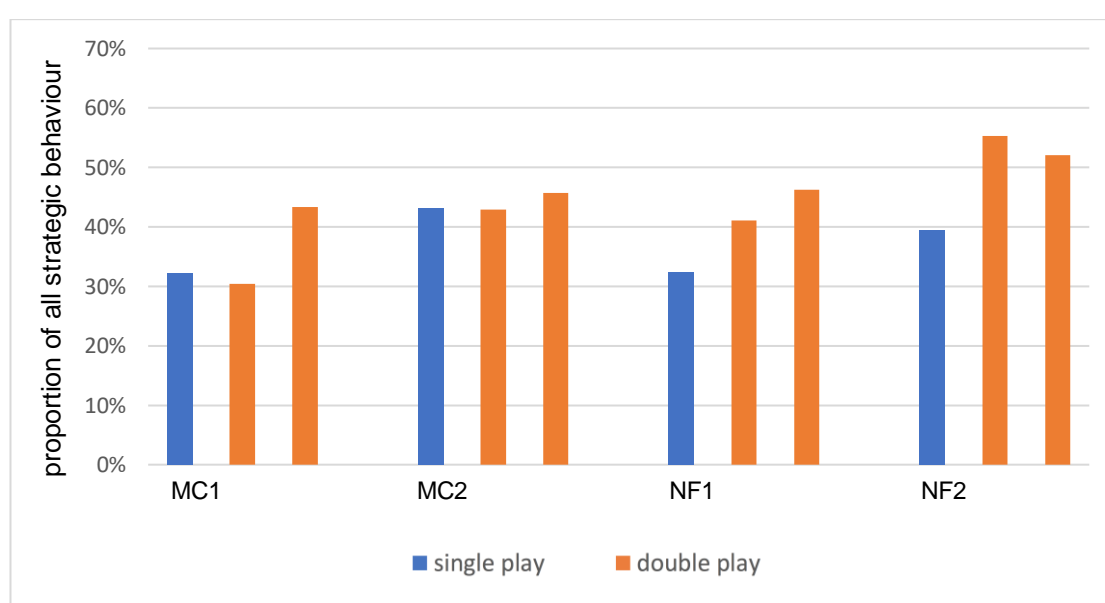
Figure 27: Study 2: total number of listening strategies as a proportion of overall metacognitive processing for single play and double play (MC and NF; by stage of task completion)



To investigate whether these task type effects are consistent with findings on task level, the data for the while-listening stage was also analysed separately for each of the four tasks. The data for the pre-listening and post-listening stages were not analysed on task level due to the small sample size of quotations associated with these stages of recall. The results are shown in Figure 28. Overall, the observation of increased levels of listening strategies only during the second play of double play versus single play for MC tasks but already during the first play of double play for NF tasks is consistent with findings on task level. However, there were differences in listening-strategic behaviour between the individual tasks. Whereas there was a marked increase in the use of listening strategies between single play and the second play of double play for MC1,

the difference for MC2 was smaller. For the NF tasks, levels of listening strategies did increase already during the first play of double play versus single play for both tasks, however, while they increased further during the second play for NF1, they dropped slightly during the second play for NF2. This is again congruent with the findings on cognitive processes, where it was shown that levels of higher-order cognitive processing dropped during the second play for NF2 compared to the first play. In addition, overall levels of listening strategies were higher for NF2 than for NF1.

Figure 28: Study 2: total number of listening strategies as a proportion of overall metacognitive processing for single play and double play (MC1, MC2, NF1, NF2; while-listening only)



Despite these differences, the general trend was the same for all four tasks: Students engaged in higher levels of listening strategies in double play compared to single play, and within double play, listening strategies were typically more prevalent during the second play than the first.

5.3. Test-taking strategies

The two different types of test-taking strategies – test-management strategies and test-wiseness strategies – were both observed in the data. As with cognitive processes and listening strategies, exemplary quotations for the two strategies will be presented first, before analysing to what extent the participants' reliance on test-taking strategies

differed between the two conditions (RQ 4), the two task types, and the four tasks (RQ 4a).

5.3.1. Examples from the data

In total, 543 quotations were coded as test-taking strategies. The vast majority of these (N=520) were related to test-management, whereas only a small number (N=23) were coded as test-wiseness strategies. Test-management was displayed by all 16 participants and test-wiseness by 12 of the 16 participants (see Appendix 11).

Test-management strategies, as defined in this thesis, are controlled and goal-directed mental actions with the ultimate goal of finding an answer to a question. For this reason, test-management strategies rely on reference to the test itself (the test questions, answer options, or chosen answers) or on expectations about how the test works. However, crucially, they are also informed by the listening text. A typical example of test-management is when a candidate attempts to answer a question based on their understanding of the relevant passage:

P03 MC1 twice while-listening first

So here I tried to decide which one of these two options I would choose, because both were mentioned, but I wasn't sure which one is correct.

In some instances, test-management was only evident in combination with either a listening strategy or a cognitive process. The following example from participant 4, which was coded as lexical search and test-management, illustrates this:

P04 NF2 once while-listening

P04: Here [...] he mentioned "social life" [...] so I wrote down "social life".

In contrast to test-management strategies, which are informed by both the test itself as well as the candidate's understanding of the listening text, **test-wiseness strategies** are either solely based on the test (or expectations about how the test works), or on some other construct-irrelevant external source such as guessing. For example, in the following quotation, stemming from the stimulated recall of participant 11, the test taker excludes two out of four answer options while reading through the test questions

before listening to the text, because they thought that these two answer options sounded more plausible than the other two options:

P11 MC1 twice pre-listening first

P11: Here I thought immediately [that] this is either “changed for the better” or “become a famous politician”.

INT: Before you even listened to the text?

P11: Before I even listened to the text.

Another typical example of test-wiseness is guessing, as the following excerpt from participant 12 illustrates:

P12 NF2 once while-listening

P12: So I didn’t know [the answer to] the first question. I thought he would talk about that later but then in the end I just guessed.

5.3.2. Analysis of test-taking strategies in single and double play

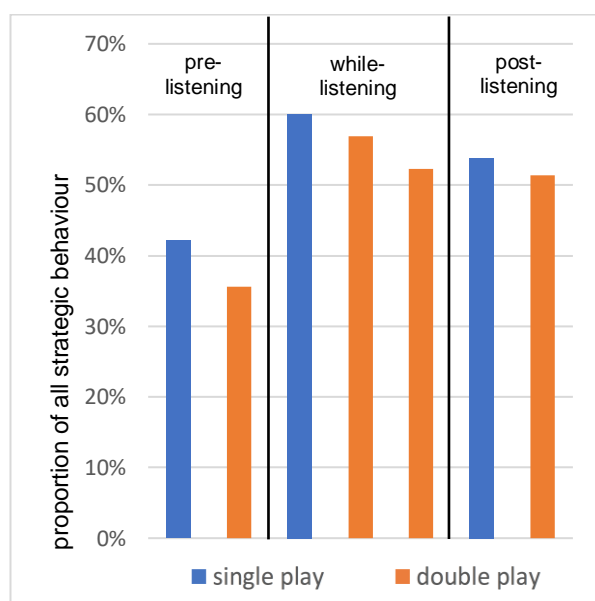
The data for test-taking strategies was analysed in the same way as the data for listening strategies. That is, the amount of test-strategic behaviour in the two conditions was calculated as a proportion of the total amount of listening strategies and test-taking strategies for each stage of task completion. First, the data was analysed for all tasks jointly (RQ 4), followed by separate analysis of the two task types and the four tasks (RQ 4a). All analyses were run twice: once for test-management strategies and once for test-wiseness strategies.

5.3.2.1. Test-management strategies

The results for test-management for all tasks jointly are shown in Figure 29. Similar to the graphs on listening strategies presented above, the chart juxtaposes the two conditions (single play in blue and double play in orange) for each stage of task completion. As shown in the figure, test-management formed a large part of students’ strategic behaviour across all stages, ranging from 36% (pre-listening in double play) to 60% (while-listening in single play). However, there was a clear tendency for each stage of recall: candidates engaged in greater relative amounts of test-management in single play as compared to double play. The difference was most striking in the pre-

listening stage (42% in single play versus 36% in double play) and in the while-listening stage (60% in single play versus 52% in the second play of double play). In the post-listening stage, the difference between the two conditions was smaller, but pointed in the same direction (54% in single play and 51% in double play).

Figure 29: Study 2: test-management strategies as a proportion of overall metacognitive processing for single play and double play (all tasks; by stage of task completion)



Thus, overall, the results suggest that candidates were more focussed on answering the test questions when they completed the tasks in a single play condition compared to a double play condition. This trend can also be illustrated with quotations from the data, such as the following example from the post-hoc interview of participant 7:

P07 post-hoc

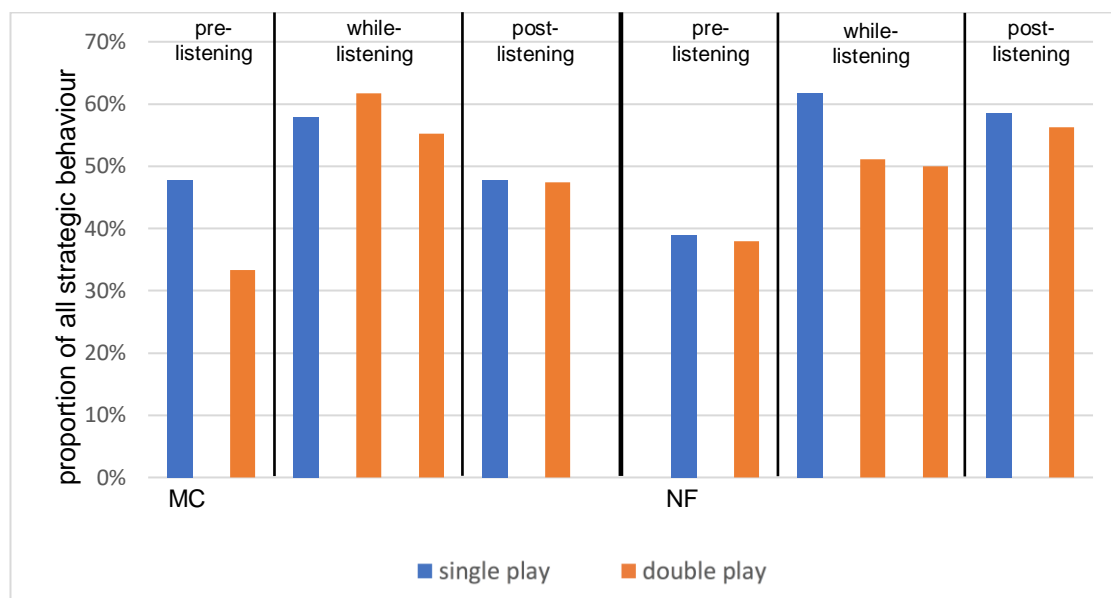
When I know that I'm only going to hear it once I try to listen more closely. [...] So it's, like, I'm listening and at the same time I'm reading through the questions and look whether they say something about the question and then I move on to the next question. When I hear it twice I can, like, focus on listening during one play and only fill in what I think might be right. [...] I can't do that when I hear it only once, because then I need to answer immediately.

To investigate in what ways the two task formats impacted the use of test-management strategies (RQ 4a), the same analysis was performed separately for each task type. The results are displayed in Figure 30. As shown in the graph, in the pre-listening stage the two task types performed in line with the overall trend: test takers relied more on test-management in single play than double play. However, the difference between the two conditions was more pronounced for the MC tasks compared to the NF tasks, indicating that during single play MC tasks students focussed more on the test paper before listening to the text than during single play NF tasks. This may have to do with the amount of text in the test paper: For MC tasks test takers not only need to read through the questions but also the four answer options, whereas in NF they only need to study the questions.

In the while-listening stage the two task types also behaved slightly differently. For the NF tasks, participants' reliance on test-management was considerably higher in single play (62%) than double play (51% in the first play and 50% in the second play). For the MC tasks, on the other hand, participants engaged in higher levels of test-management during the first play of the double play condition (62%) than the single play condition (58%) but in lower levels during the second play (55%).

Finally, differences between the two task types were also found in the post-listening stage. Levels of test management were markedly higher for the NF tasks than the MC tasks, which might again have to do with the nature of the task format: In NF, students used the time after the recording finished playing to fine-tune their open answers, whereas in MC they simply checked their answers. In addition, there was no marked difference in test-management between single play and double play for the MC tasks, whereas for the NF tasks the general trend continued: Students engaged in higher levels of test-management in single play than double play.

Figure 30: Study 2: test-management strategies as a proportion of overall metacognitive processing for single play and double play (MC and NF; by stage of task completion)



Despite these differences, the general trend of increased levels of test-management during single play versus double play was similar for both task types. The following examples from the data illustrate this. The quotations are taken from the retrospective recall of participant 15 after they experienced MC1 in single play and from the stimulated recall of participant 14 while they completed NF2 in single play:

P15 MC1 once retrospective

Yes, so now [in single play] I looked very closely at the questions, because, like, as I said before, [during double play] I first listen [...], but that would not be so wise now, because I won't hear it a second time.

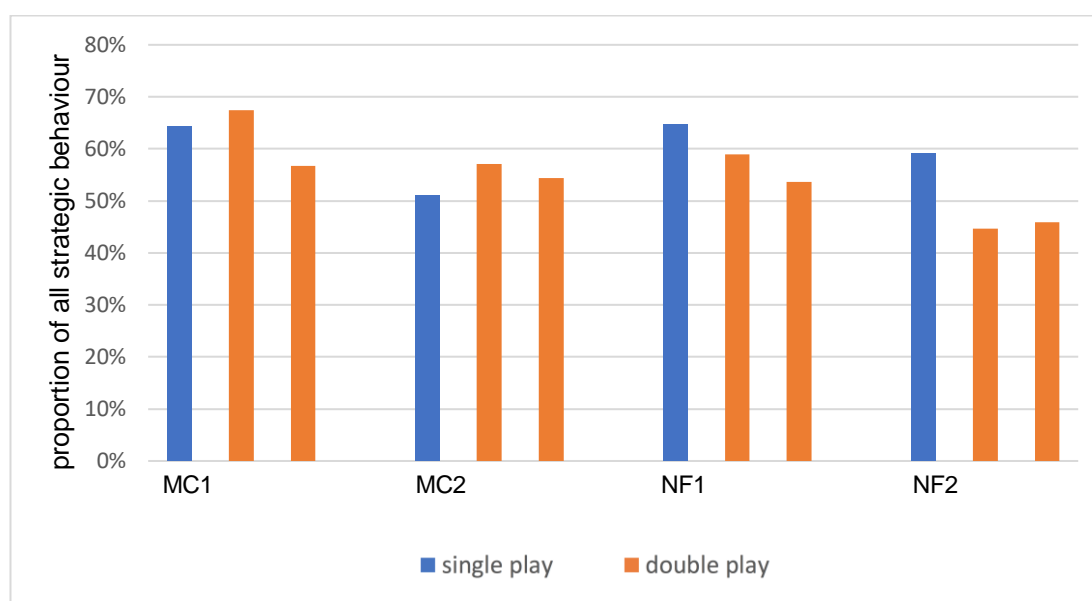
P14 NF2 once pre-listening

Yes so [when I saw that I would hear it only once] I suddenly realised that it's going to be a little more difficult now. So I realised [...] that I would need to prepare more in terms of reading [the questions].

As a last step the data for the four tasks was analysed separately to discern whether individual task effects are in line with the findings on the two task types. For this analysis, only the data for the while-listening period was used, as the sample size of quotations for the pre-listening and post-listening stages at task level was too small to

allow for meaningful comparisons. The results are presented in Figure 31. It can be seen that both MC tasks display a similar pattern, with increased levels of test-management in the first play of double play compared to single play and a subsequent drop during the second play. However, while levels of test-management during the second play decreased markedly below those of single play for MC1, they remained slightly higher during both plays of double play compared to single play for MC2. A different pattern was observed for the NF tasks, where levels of test-management were already lower during the first play of double play compared to single play. For NF1, the relative amount of test-management decreased even further during the second play, while it increased again slightly for NF2. It can also be seen that the difference between single play and double play was larger for the NF tasks than the MC tasks. Thus, overall, these results seem to suggest that levels of test-management during the while-listening period are impacted more in NF tasks than MC tasks.

Figure 31: Study 2: test-management strategies as a proportion of overall metacognitive processing for single play and double play (MC1, MC2, NF1, NF2; while-listening only)

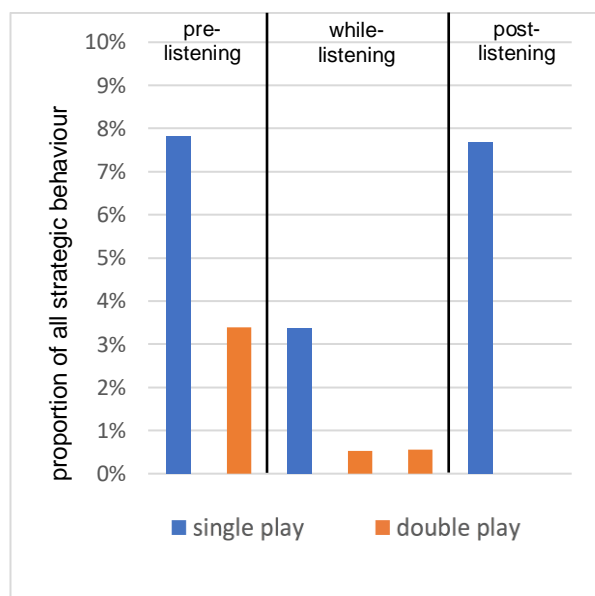


5.3.2.2. Test-wiseness strategies

Despite the small number of quotations related to test-wiseness strategies, the results show a similar trend to the findings on test-management, as shown in Figure 32. In the pre-listening stage of the single play condition, participants on average spent 8% of their overall metacognitive processing on test-wiseness, compared to 3% in double play. Similarly, in the while-listening period the proportion of test-wiseness was 3% in single

play and less than 1% in double play. Finally, in the post-listening stage of the single play condition, 8% of candidates' metacognitive processing was coded as test-wiseness, but no evidence for test-wiseness was found in double play.

Figure 32: Study 2: test-wiseness strategies as a proportion of overall metacognitive processing for single play and double play (all tasks; by stage of task completion)



Overall, these results show that participants engaged in higher levels of test-wiseness strategies in single play as compared to double play. The following quotation from the post-hoc interview of participant 13 is a typical example:

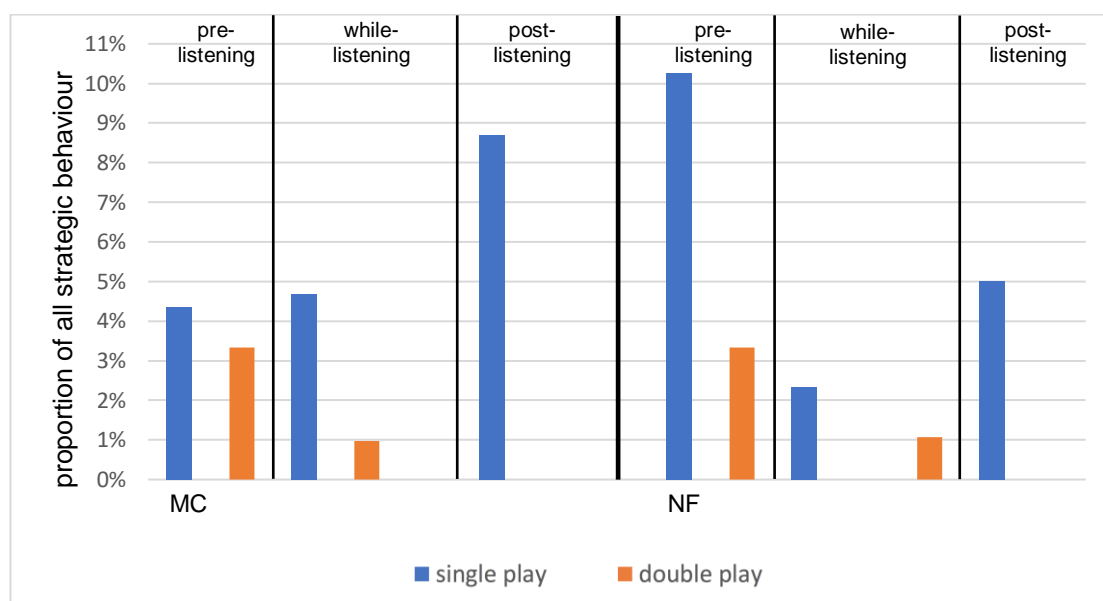
P13 post-hoc

So I think when I know that I will hear it only once [...] I try to get all [the answers] during the first play. And if I don't get something then I simply choose whatever I think fits best.

When looking at the results of the two task types individually, the same trend can be observed, with slight difference between MC and NF (Figure 33). For both task types, the majority of test-wiseness behaviour is evident in the stages before and after listening to the recording. For NF tasks in particular, students tried to come up with answers to questions before they listened to the recording. In the post-listening stage for both task types, students took a guess whenever they were uncertain about answers after listening to the recording. However, this was only the case for the single play condition.

Overall, single play seems to be more affected by test-wiseness behaviour than double play, regardless of test format. The analysis was not performed for each task individually, as the small sample size of quotations would not have allowed for meaningful comparisons.

Figure 33: Study 2: test-wiseness strategies as a proportion of overall metacognitive processing for single play and double play (MC and NF; by stage of task completion)



5.4. Anxiety

In this section, exemplary quotations coded as evidence for anxiety will be presented first, followed by an analysis of how anxious participants were in single play compared to double play (RQ 5) and, further, whether task type affected anxiety levels (RQ 5a).

5.4.1. Examples from the data

Anxiety, as used in this thesis, refers to worries, stress, or concerns in relation to listening in a foreign language specifically (Horwitz, 2010), or to evaluative situations more generally (Cassady & Johnson, 2002). Anxiety was observed for 14 of the 16 participants, with 34 quotations overall (see Appendix 11). The following example is taken from the pre-listening stage of participant 4 before they completed NF2 in single play:

P04 NF2 once pre-listening

P04: Yes, I noticed this (the participant is referring to the information that the recording would only be played once). *And I was a little/this intimidated me slightly I have to say.*

INT: Ok. Why?

P04: Because [...] I put myself under much more pressure right away, as I had the feeling that I needed to understand everything the first time already.

Another typical example is the following excerpt from participant 16. The quotation stems from the while-listening stage for NF1, which the participant completed in single play:

P16 NF1 once while-listening

INT: Ok. Is there anything else you were thinking [while completing the task]?

P16: Yes I felt stressed, because I could only hear it once.

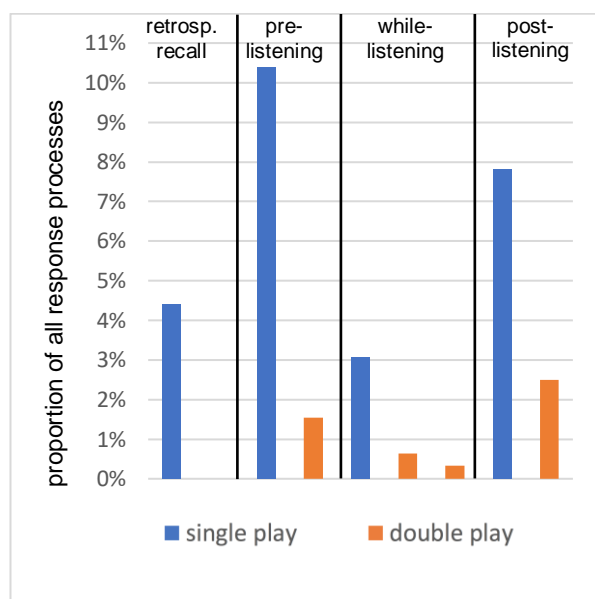
5.4.2. Analysis of anxiety in single and double play

In order to compare anxiety levels between the two conditions and across the different stages of task completion, the number of quotations coded as anxiety were divided by the total number of quotations for each stage of task completion. The analysis was first performed for all tasks jointly to answer RQ 5 and then for the two task types separately to answer RQ 5a.

Figure 34 shows the results for all tasks jointly. As with the figures on listening strategies and test-taking strategies presented above, the graph displays the two conditions in a different colour (single play in blue and double play in orange), separately for each stage of task completion. In addition to the three stages of task completion, the graph also includes data from the retrospective recalls, as participants already indicated levels of anxiety during this stage. As shown in the figure, for the retrospective recalls, 4% of all quotations were coded as anxiety in single play, but no evidence for anxiety was found in double play. In the pre-listening stage, students referred to being anxious in 10% of all quotations in single play, compared to only 2% in double play. Next, in the task-processing stage levels of anxiety dropped to 3% in single play, to 1% in the first play of double play, and close to 0% in the second play of

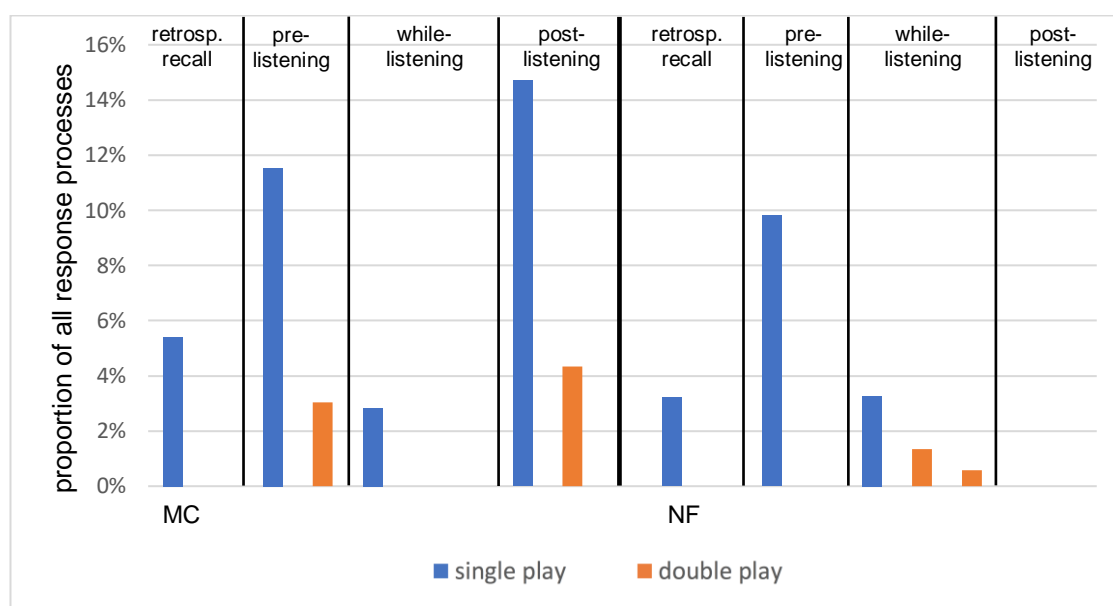
double play. Finally, in the post-listening stage, 8% of all quotations were related to anxiety in single play and 3% in double play. Thus, overall, these results show that students were markedly more anxious in single play versus double play across all stages of task completion. The results also show that anxiety levels were highest in the pre- and post-listening stages, but lower in the while-listening stage.

Figure 34: Study 2: anxiety as a proportion of all coded quotations for single play and double play (all tasks; by stage of task completion, including retrospective recalls)



In order to study potential differences in anxiety levels between MC and NF tasks, the same analysis was performed separately for the two task types (Figure 35). The results show that participants were more anxious during MC tasks than NF tasks. The difference is most striking during the post-listening stage, where test takers indicated being anxious in 15% of all quotations in single play and in 4% during double play for the MC tasks, but no evidence for anxiety was found during either condition for the NF tasks. Despite this difference, for each task type and each stage of recall within the task types anxiety levels were considerably higher in single play versus double play.

Figure 35: Study 2: anxiety as a proportion of all coded quotations for single play and double play (MC and NF; by stage of task completion, including retrospective recalls)



It was also evident from a number of quotations that anxiety negatively impacted participants' performance. The following example, taken from the post-hoc interview of participant 10, illustrates this phenomenon:

P10 post-hoc

[...] It's, like, I can't concentrate when I'm stressed and I constantly think about the stress when I hear it only once. When I know, ok, I'll hear this only once, I have to understand everything and answer everything immediately. Then I constantly think about that and can't concentrate so well.

5.5. Meta-commentary

Three different codes were applied to meta-commentary which appeared to be relevant for answering the research questions: explicit comments showing that participants displayed different listening behaviour in single play versus double play or in the first versus the second play of double play (different listening), comments indicating a preference for double play versus single play (prefer double play), and comments showing that the research method inhibited natural processing (reactivity). In the following, exemplary quotations for these three categories will be presented, followed by an analysis of how the coding categories were applied.

5.5.1. Examples from the data

Overall, 59 quotations were coded as **different listening**. In 27 of these quotations, participants clearly indicated that they listened differently when they knew that they would hear the text only once compared to when they knew that they would hear the text twice. The following example stems from the retrospective recall of participant 16 after they completed NF1 in single play (they had already completed NF2 in double play before that):

P16 NF1 once retrospective

Yes [now I was listening] in more detail, because I could only listen to it once.

In the remaining 32 quotations participants specified different listening behaviour between the first and the second play of double play, as the following quotation from participant 5 illustrates. The quotation was taken from the while-listening stage of the second play in double play of MC1:

P05 MC1 twice while-listening second

I was focusing more on what he was saying. The first time I was more, like, reading the questions.[...] But now I understood more, I was more concentrated.

The second category of meta-commentary was **prefer double play**. This code was applied to quotations where participants stated that they prefer the double play condition compared to the single play condition. In total, 24 quotations were assigned this code. Out of these, 16 quotations emerged during the post-hoc interview, as one question specifically asked whether participants preferred single play or double play:

P06 post-hoc

INT: And what do you prefer, listening once or listening twice?

P06: Listening twice.

The remaining 8 quotations stemmed from the other stages of recall, such as the following example from the pre-listening stage of participant 3 while they completed MC1 in double play:

P03 MC1 twice pre-listening first

[...] And then I saw that it [will be played] twice now. [...] I prefer it when I can hear it twice.

The third category of meta-commentary which emerged during the coding was **reactivity**. Reactivity was assigned to passages in which participants indicated that the research method inhibited their natural processing. In total, only 6 quotations were assigned this code. The following is a typical example:

P07 MC twice retrospective

So it was a little different [to a normal exam situation] because I had to concentrate on the computer [screen]. [...] And it irritated me a little that I could not move my head [...] because normally I would look around in class.

5.5.2. Analysis of meta-commentary

The three categories of meta-commentary were analysed in terms of how many of the participants mentioned the categories in their recalls. For the category different listening, 14 of the 16 candidates stated that they listen differently if they know from the outset that they would hear the recording only once compared to when they know they would hear it twice. In terms of different listening behaviour between the first and second play of double play, all of the candidates mentioned this at least once in their recalls. These results thus confirm the findings on the four response processes of interest presented above, where differences between the two conditions were found for cognitive processes, listening strategies, test-taking strategies and anxiety levels.

The second category of meta-commentary, prefer double play, was observed for all participants. That is, all 16 participants clearly indicated that they prefer double play to single play. This is not surprising as all participants were used to double play, as it is currently the norm in the Austrian Matura context. Participant 8 stated that they would prefer double play even if the listening task was in German:

P08 post-hoc

INT: What do you prefer, listening once or listening twice?

P08: Definitely listening twice. [...] If the listening was in German, I would also prefer listening twice.

Finally, only 4 out of 16 participants referred to reactivity effects during their recalls; i.e. they mentioned that the research methodology inhibited their natural processing. All four participants found it distracting that they were not allowed to move their head during the experiment and two of them indicated that they were not used to taking listening tests on a computer. However, despite these reactivity effects, when asked about the research methodology in the post-hoc interview, 15 of the 16 candidates stated that seeing their eye-movements helped them remember what they were thinking during the test:

P01 post-hoc

INT: How did you find watching the video of your eye-movements? Useful or distracting?

P01: It was very useful. When I looked at a word for some time, then it was, like: "Sure here I was thinking this, I remember again". Only seeing my answers again would not have helped me nearly as much.

5.6. Summary of the main findings

Overall, it can be concluded that construct-related differences existed between the single play and double play conditions with regards to all four response processes of interest. Participants displayed a larger amount of higher-level cognitive processes (meaning construction and discourse construction) and a smaller amount of lower-level cognitive processes (lexical search and parsing) in double play compared to single play. Participants also engaged in a greater variety of listening strategies, and the use of all but two listening strategies was higher in double play than single play. In addition, participants used fewer relative amounts of test-management and test-wiseness strategies and they were markedly less anxious. Within the double play condition, these effects were typically stronger during the second play than the first. The results were confirmed by an analysis of relevant meta-commentary, where it was shown that students listened differently depending on the condition. Students also clearly preferred double play over single play. In sum, these results confirm the findings of Study 1.

Although overall findings are consistent with the results on the individual task types, a number of differences between MC and NF with regards to the four response processes were observed in Study 2. First, for MC tasks increased levels of higher-order cognitive processing and listening-strategic behaviour were only evident during the second play of double play compared to single play, but not during the first. For NF tasks, on the other hand, the shift to increased higher-order cognitive processing and listening-strategic behaviour was generally already prevalent during the first play of the double condition, and for NF1 levels increased further during the second play. In addition, test takers in general displayed more higher-order cognitive processing for MC tasks, which was to be expected as the MC items targeted main ideas and supporting details whereas the NF tasks targeted specific information and important details. Second, levels of test-management strategies were higher for MC tasks than NF tasks in the pre-listening stage for single play, but higher for NF tasks than MC tasks in the post-listening stage for both conditions. In the while-listening stage, test-management strategies decreased more in NF tasks than MC tasks in double play versus single play, particularly during the second play of the double play condition. Finally, despite a strong decrease of anxiety in double play compared to single play for both task types, it was shown that participants were generally more anxious during MC tasks than NF tasks, most notably during the post-listening stage.

6. Discussion

The discussion of the findings is divided into four main sections. First, the results of the two individual studies will be converged and summarised in Section 6.1 and it will be discussed to what extent the two separate studies agreed in answering the five research questions. Second, in Section 6.2 the results will be compared to past research on double play in L2 listening assessment, with a particular focus on the only previous large-scale study on double play by Field (2015). Section 6.3 then extends our current theory of listening assessment in light of the research in this thesis by considering the impact of single play versus double play on the construct that is measured. Finally, in Section 6.4 the findings of the two studies are discussed with regards to competing priorities in listening assessment, including test purpose, cognitive demand, reliability, and practicality.

6.1. Convergence and summary of findings

Before discussing how the results relate to previous research and how they extend our current theories of listening assessment, in this section the findings of the two studies will be converged and summarised. The structure of the section follows the research questions. Table 38 shows an overview of the research questions and the types of data used to answer them.

Table 38: Types of data used to answer the research questions

Research question	Types of data
RQ 1: item and task statistics	test scores after single play (<i>Study 1</i>) test scores after the first and second play of double play (<i>Study 1</i>) Questionnaire 2 (<i>Study 1</i>)
RQ 2: cognitive processes	Questionnaire 2 (<i>Study 1</i>) verbal recalls (<i>Study 2</i>)
RQ 3: listening strategies	Questionnaire 1 (<i>Study 1</i>) answer changes (<i>Study 1</i>) verbal recalls (<i>Study 2</i>)
RQ 4: test-taking strategies	Questionnaire 1 (<i>Study 1</i>) answer changes (<i>Study 1</i>) verbal recalls (<i>Study 2</i>)
RQ 5: anxiety	Questionnaire 1 (<i>Study 1</i>) verbal recalls (<i>Study 2</i>)

6.1.1. Item and task statistics in single and double play

The first research question was:

1. What are the differences in item and task statistics between single and double play listening tasks?
 - a. Is task type a factor?

This research question was answered in Study 1, where 306 test takers completed four listening tasks in two formats (MC and NF) across single play and double play in a complex and carefully counterbalanced research design. Two main findings emerged. First, the results of a CTT analysis revealed that overall reliability and item discrimination are enhanced by the double play condition compared to the single play condition (RQ 1). This effect was similar between the two task formats but it was slightly larger for MC tasks than NF tasks, indicating that double play improves item properties more for MC tasks than for NF task (RQ 1a). Overall, however, item and task statistics were better for NF tasks than MC tasks across both conditions.

Second, an MFRM bias analysis showed that test items are significantly easier in double play versus single play (RQ 1), with a difference in average item difficulty of .93 logits for MC tasks and .55 logits for NF tasks. However, significantly higher test scores were also observed in single play compared to the first play of double play (.70 logits difference for MC tasks and .44 logits for NF tasks), and in the second play of double play compared to the first (1.49 logits difference for MC tasks and 1.00 logits for NF tasks). All of the effects were larger for MC tasks than NF tasks, suggesting that MC tasks benefit more from a second play in terms of increased test scores than do NF tasks (RQ 1a). However, overall the difference in task difficulty between MC and NF tasks was smaller in double play than single play, indicating that the two task types are more comparable in the double play condition.

These results were confirmed by an analysis of responses to Questionnaire 2, where one item asked students to indicate on a 4-point Likert scale how difficult they found the listening tasks. Mann-Whitney U tests were calculated to explore differences in perceived task difficulty between the two conditions and task types within each sub-group of students. The results showed that participants from both sub-groups perceived the tasks to be significantly easier in double play compared to single play. Effect sizes were medium to large for sub-group 1 and small to medium for sub-group 2.

6.1.2. Cognitive processes in single and double play

The second research question was:

2. What are the differences in test takers' cognitive processes between single and double play listening tasks?
 - a. Is task type a factor?

This research question was informed by both studies but was mainly answered in Study 2, where 16 candidates completed the same tasks as in Study 1 in single and double play and eye-tracking was used in combination with verbal recall to gain insights into candidates' response processes. Evidence for four cognitive processes was found in the verbal report data: lexical search and parsing (lower-order processes) and meaning construction and discourse construction (higher-order processes). The data analysis showed that in both conditions the majority of observed cognitive processes

were related to lexical search and parsing. However, there was a clear difference in cognitive processing between single play, the first play of double play, and the second play of double play: Overall, participants displayed a larger amount of higher-order cognitive processes (meaning construction and discourse construction) and a smaller amount of lower-order cognitive processes (lexical search and parsing) in both plays of the double play condition compared to the single play condition. Within double play, increased levels of higher-order cognitive processing were observed in the second play versus the first. The findings were substantiated with direct quotations from the verbal reports. Overall, these results suggest that test takers engaged more in meaning-building processes in double play as compared to single play, particularly during the second play of the double play condition (RQ 2).

To some extent these results were confirmed in Study 1 by an analysis of responses to Questionnaire 2, where one item asked how well students were able to show their listening proficiency through the tasks used in the study (perceived validity). Both sub-groups of students felt that they were better able to demonstrate their listening proficiency in double play than single play, with medium to large effect sizes. Although this is not direct evidence for increased levels of higher-order cognitive processes, it indicates that students may have used a greater variety of cognitive listening processes in double play compared to single play.

In terms of task type effects (RQ 2a), no clear differences between MC and NF were observed with regards to the single and double play convention. For three of the four tasks used in the research, students were generally listening more locally by displaying a larger amount of lower-order cognitive processes during single play and the first play of double play, but more globally by using increased levels of higher-order cognitive processes during the second play of double play. Only for NF2 the increase in higher-order cognitive processing between single and double play was larger during the first play of the double play condition. During the second play test takers fell back on lower-order cognitive processing, albeit still to a lesser extent than during single play. One possible explanation for this difference is the fact that NF2 was based on a shorter listening text and was somewhat more difficult than the other three tasks, which might have impacted test takers' response processes. In general, however, test takers displayed more higher-order cognitive processing for MC tasks compared to NF tasks, which was to be expected as the MC items targeted main ideas and supporting details whereas the NF items targeted specific information and important details.

6.1.3. Listening strategies in single and double play

The third research question was:

3. What are the differences in test takers' use of listening strategies between single and double play listening tasks?
 - a. Is task type a factor?

Both studies provided data on RQ 3. In Study 1, candidates indicated their use of listening strategies by choosing their level agreement to a number of statements in Questionnaire 1. Candidates filled out the questionnaire twice – once after they had completed two tasks (of the same format) in a single play condition and once after they had completed the other two tasks (of the other format) in a double play condition. In addition, the use of listening strategies could be inferred from a detailed analysis of answer changes during the second play of double play. In Study 2, the stimulated recall data was analysed in terms of the relative number of listening strategies used by participants in single play, the first play of double play, and the second play of double play across the different stages of task completion.

The findings of the two individual studies show the same trend. In Study 1, the results for item difficulty show that first listening on double play is not equivalent to single play, which suggests that the first listening on double play is used for different strategic purposes. This was confirmed by a Wilcoxon signed-rank test based on an exploratory factor analysis of questionnaire items, which revealed that candidates used significantly more listening strategies in double play compared to single play, with a small effect size. In addition, the analysis of answer changes in the second play of double play showed that participants used the second play to revise a substantive proportion of their answers in a variety of ways and, in the case of NF tasks, to add more details or choose a different correct answer. These findings indicate that participants understood more of the listening text during the second play, or at least that they had more opportunities to showcase their understanding. These results were confirmed in Study 2, where I showed that participants engaged in a greater variety of listening strategies in double play than single play. In addition, the proportion of listening strategic behaviour compared to overall metacognitive processing in the while-listening stage was higher in double play than single play for 8 out of 10 observed

listening strategies. This effect was strongest in the second play of the double play condition. Thus, overall, the results show that participants engaged more with the listening text during double play versus single play by displaying a greater variety and a greater relative amount of listening-strategic behaviour.

In terms of task type effects (RQ 3a), the stimulated recall analysis showed that although the general trend of increased levels of listening strategies in double play versus single play was found across all stages of task completion for both task types, there was a difference between the two task types in the while-listening stage. For MC tasks, increased levels of listening-strategic behaviour were only found for the second play of double play compared to single play, but not for the first play. For NF tasks, on the other hand, candidates' use of listening strategies was markedly higher already during the first play of double play compared to single play, and, in the case of NF1, increased further during the second play.

6.1.4. Test-taking strategies in single and double play

The fourth research question was:

4. What are the differences in test takers' use of test-taking strategies between single and double play listening tasks?
 - b. Is task type a factor?

Similar to RQ 3, RQ 4 was answered through data collected in both studies. In Study 1, candidates indicated their use of test-taking strategies by completing Questionnaire 1 after the single play condition and again after the double play condition. In Study 2, I calculated the amount of test-strategic behaviour in single play, the first play of double play, and the second play of double play as a proportion of all metacognitive processing across all stages of task completion. Following the definition of test-taking strategies by A. D. Cohen (2011), I differentiated between test-management strategies and test-wiseness strategies in the analysis of the verbal report data.

As with the results on listening strategies, the findings of the two studies agree. Candidates used more test-taking strategies in single play compared to double play. In Study 1, this was demonstrated through a Wilcoxon signed-rank test based on an

exploratory factor analysis of questionnaire items, which revealed a statistically significant result with a small effect size. In addition, in the open question of Questionnaire 1, 68 participants explicitly stated that in single play they struggled with the multitasking demands of the listening test (simultaneous listening to the text, reading the questions, thinking about the answers, writing down the answers, and checking the answers). Concordantly, in Study 2 the analysis of the verbal report data clearly showed that in the great majority of tasks and stages of task completion candidates were less reliant on test-taking strategies in double play versus single play, particularly during the second play of the double play condition, both in terms of test-management strategies and test-wiseness strategies.

In terms of task type effects (RQ 4a), the stimulated recall analysis showed that MC and NF tasks seem to be impacted slightly differently by the double play condition with regards to the use of test-taking strategies. In the pre-listening stage, the drop in test-management strategies in double play compared to single play was more pronounced for MC tasks compared to NF tasks, whereas test-wiseness dropped more markedly for NF than MC. Also, in the while-listening stage, levels of test-management were impacted more in NF tasks than MC tasks in double play. Finally, in the post-listening stage, levels of test management were markedly higher for NF tasks than MC tasks. Despite these differences, the overall trend was the same for both task types, in that students engaged in higher levels of test-management and test-wiseness in single play compared to double play.

However, the analysis of answer changes during the second play of the double play condition in Study 1 indicates that in MC tasks candidates may be more likely to fall back on test-wiseness than in NF tasks. After the second listening, a total of 5.9 percent of all NF answers were left blank, compared to only 0.5 percent of all MC answers. The difference was statistically significant with a large effect size. This seems to be evidence that MC tasks are more prone to guessing than NF tasks.

6.1.5. Anxiety in single and double play

The fifth research question was:

5. What are the differences in test takers' anxiety levels between single and double play listening tasks?
 - c. Is task type a factor?

This research question was answered by referring to data from both studies. In Study 1, the analysis of responses to Questionnaire 1 showed that students were significantly more anxious in single play compared to double play, with a small to medium effect size (RQ 5). In addition, I identified 108 comments relating to increased anxiety in single play in the analysis of responses to the open question of Questionnaire 2. In these comments, the students referred to feeling panicked, being stressed, under pressure, or nervous, which either inhibited their listening ability or made them feel insecure or uncomfortable. These results were confirmed in Study 2, where the analysis of the verbal report data clearly showed that students were markedly more anxious in the single play condition versus the double play condition and that anxiety negatively impacted students' performance. As with the findings on RQ 2, RQ 3, and RQ 4, I substantiated the frequency distributions with quotations from the verbal reports.

The verbal report data also displayed a difference in students' anxiety levels between MC and NF tasks (RQ 5a). It was shown that students were generally more anxious in MC tasks compared to NF tasks, particularly during the pre- and post-listening stages. Despite this difference, anxiety levels were considerably higher in single play versus double play for both task types across all stages of recall.

6.2. Connection with previous research

In this section the results will be discussed in light of past research on double play in L2 listening assessment. First, the findings on RQ 1 will be connected to earlier studies on this topic, most of which investigated whether double play had any effect on test scores and item statistics. Second, particular attention will be given to the study by Field (2015), which is the only previous large-scale investigation of test takers' response processes in relation to double play.

The great majority of earlier studies on the effects of double play found that it aided comprehension and increased test takers' scores (Berne, 1995; Chang & Read, 2006; Field, 2015; Iimura, 2007; Lund, 1991; Ruhm et al., 2016; Sakai, 2009), while a smaller number of studies reported that students did not benefit from double play as much as expected (Brindley & Slatyer, 2002; Henning, 1991). However, most of these earlier studies looked at double play as a secondary treatment as part of a larger investigation, while only four studies focussed exclusively on double play (Field, 2015; Iimura, 2007; Ruhm et al., 2016; Sakai, 2009). In addition, apart from Brindley and Slatyer's (2002) research, none of the studies investigated the effects of single and double play in a counter-balanced design, but either compared different tasks and test takers across the two conditions or only contrasted the first and second play of a double play condition. My study is unique in that it directly compared the effects of single and double play in a complex counter-balanced experimental design, utilizing listening tasks which had been professionally developed according to state-of-the-art international standards including piloting and standard setting, and controlling for confounding factors such as task format, targeted level, task difficulty, number of items per task, topics covered by the tasks, sound file length, targeted listening behaviour, and task ordering effects.

The results of Study 1 suggest that double play increases test scores, item discrimination, and overall reliability, which agrees with the main share of previous research but is contrary to the findings by Brindley and Slatyer (2002) and Henning (1991). However, a number of factors might have impacted the findings of these two studies. First, the studies by Brindley and Slatyer (2002) and Henning (1991) did not look at double play alone, but also investigated other variables such as speech rate, text type, live versus recorded materials (Brindley & Slatyer, 2002), length of listening text and associated number of items, reading response length, and level of processing skills (Henning, 1991), thereby necessarily limiting both the complexity and thoroughness of the research design with regards to double play, as well as the number of participants for each condition. For example, in Henning's study only about 40 participants completed each of the tasks in double play. Also, the listening tasks used in the studies by Brindley and Slatyer and Henning may not have been ideal. Brindley and Slatyer based their research on tasks developed by teachers within an adult migrant English program without formal training in language assessment, while Henning used tasks consisting of isolated one-, two-, or three-sentence listening passages taken from

TOEFL tests from the mid-1980s. Taking these limitations into account and considering that the great majority of previous studies agree with my finding on increased test scores in double play, it seems reasonable to conclude that double play increases test scores and is also beneficial with regards to overall reliability and item discrimination.

At first sight it might seem counter-intuitive that double play increases discrimination and reliability, as one could assume that differences between test-takers may be evened out through the double play condition. However, this position does not take construct-irrelevant factors into account. I have shown that construct-irrelevance – particularly test-wiseness strategies such as guessing and feelings of anxiety – is minimised in double play. This arguably leads to a more accurate reflection of true scores and therefore a better overall relationship between item and total score in the item discrimination analysis. This, in turn, would also explain the better reliability figures for the tests in the double play condition. Thus, overall, the improved item statistics may be an indication that construct-irrelevant factors such as test-taking strategies and anxiety are minimised under double play compared to single play.

Another important finding of my research with regards to task difficulty is that selective response items (MC tasks) benefitted more from the second play than constructed response items (NF tasks) in terms of increased test scores. This is contrary to results from Field (2015), who found a significantly higher increase in test scores in the second play of double play for constructed response items compared to selective response items. While one possible explanation for this discrepancy could be the relatively small number of participants in Field's study, which may not allow for generalizations (he compared groups of 33 and 40 students), the results might also have been impacted by the nature of the tasks used. Field used a three-option MC format, but the tasks used in my study were four-option MC, resulting in a higher reading load for participants. Similarly, in the constructed response tasks of my study candidates had to complete one gap at the end of sentences, whereas Field used items with gaps in the middle of sentences, which makes the items cognitively more demanding as students have to formulate their answers to match both the beginning and the end of the sentence. Thus, in my study students may have benefitted more from the second play for the MC items due to the high reading load of these items, but less for the relatively straightforward NF items, whereas students in Field's study may have profited more from the second play for the cognitively demanding gap-fill items than the comparatively shorter MC items (see also Kintsch, 1998). Despite this discrepancy, my

results agree with Field (2015) in that different task formats are more comparable in terms of task difficulty in double play versus single play, that is the task format impacted item difficulty less in double play compared to single play. This finding also aligns with the argument outlined above that method effects (e.g. construct-irrelevant variance) are minimised under a double play condition.

The effects of double play on test scores and statistical item parameters are worth addressing, but for making decisions about language test design it is arguably more relevant to ask whether double play impacts the construct that is measured. As pointed out by Messick (1995), in order to investigate construct validity “possibly most illuminating of all [...] are direct probes and modeling of the processes underlying test responses” (Messick, 1995, p. 743). Field’s (2015) study is the only previous large-scale investigation that has looked at response processes in relation to double play in listening assessment. Contrary to my study, though, Field (2015) did not compare test takers’ response processes between single play and double play, but only between the first and second play of a double play condition.

Field (2015) found that during both the first and second play in double play the majority of participants relied on lower-order cognitive processes, and many participants used higher-order cognitive processes only during the second play. My findings from Study 2 show the same trend. However, due to my research design I was able to detect greater reliance on lower-order cognitive processes not only for the first play of double play compared to the second play, but also for single play compared to double play. In addition, I showed that for one of the NF tasks students made use of increased levels of higher-order cognitive processes already during the first play of double play compared to single play.

Two other areas where Field’s (2015) and my findings conform are anxiety and test-wiseness, but because of the more comprehensive research design of my study I was able to refine and extend upon Field’s results. Both of our studies found increased levels of anxiety and test-wiseness during the first play of double play compared to the second play. However, importantly, I was able to show that levels of anxiety and test-wiseness strategies are markedly higher in single play than in either play of the double play condition. In addition, students displayed the highest levels of anxiety and test-wiseness in the stages before and after listening to the texts, but were considerably less anxious and less reliant on test-wiseness strategies in the while-listening stage. Field

only reported on the while-listening stage of the first and second play in double play and was thus not able to detect these effects.

My study also shed light on the impact of double play on candidates' use of test-management and listening strategies - two important types of response processes which Field (2015) did not investigate. I showed that candidates were less reliant on test-management strategies and that they used a greater variety and greater relative amount of listening strategies in double play compared to single play. These effects were particularly prevalent during the while-listening stage in the second play of the double play condition, but were consistent throughout the different stages of task completion. In addition, students relied more on specific listening strategies during the first play of double play (prediction and monitoring), whereas for others, indeed for the majority, they used higher levels during the second play (focusing attention, evaluation, inferencing, elaboration, prediction, contextualization, translation, and managing emotions). This is further evidence that the construct-beneficial effects of double play do not exclusively manifest themselves only during the second play of double play.

Another area which was investigated by both Field (2015) and in my research was the nature of answer change during the second play. Field found that for about 50 percent of all responses during the second play candidates left an incorrect answer unchanged, i.e. they did not seem to benefit from the second play in terms of increased understanding. In my study this number was much lower, with only about 13 percent of all responses being incorrect after the first play and still incorrect after the second play. Similarly, in Field's study only about 13 percent of all responses were changed from incorrect after the first play to correct after the second play, whereas in my study beneficial answer changes amounted to 26 percent for MC items and 22 percent for NF items. However, crucially, participants in Field's quantitative study did not know from the beginning that they would hear the recording a second time, but were only told after the first play that they would get a second chance. Participants in my study, in contrast, knew from the start that they would hear the recording twice, as they would also be provided with this information in a real-life listening test. Thus, these different findings seem to be further evidence that double play is not simply a repetition of a single play condition, but rather that test takers are behaving differently from the beginning if they know that they will hear the text a second time.

In sum, the research findings in this thesis both support and challenge the existing knowledge base concerning repetition of listening test input, and in a number of

important aspects novel and more nuanced findings emerged. In terms of item properties, the results confirm the main share of previous findings in that double play increases tests scores. I also showed that double play is beneficial for reliability and item discrimination, which is contrary to results by Brindley and Slatyer (2002) and Henning (1991). However, it was argued that the research design of these particular studies may not have been ideal for investigating the effects of single play versus double play. Also, the research findings confirm the results by Field (2015), but because of the complex counter-balanced research design of my study I was able to build on Field's findings and address previously unexplored aspects. My study is the first to directly compare single play with double play and the first to systematically investigate how either condition affects test takers' use of listening and test-taking strategies as well as their anxiety levels.

6.3. Extending the theory of listening assessment

The research presented in this thesis provides important insights on the effects of double play in L2 listening assessment and by so doing extends our understanding of the listening construct. The study addresses the urgent need of language assessment practitioners to more fully understand “the mechanisms that underlie what people do, think, or feel when interacting with, and responding to, [listening items or tasks] and are responsible for generating observed test score variation” (Hubley & Zumbo, 2017, p. 2). This is a pressing issue, since “from a primarily cognitive perspective, the processes involved in second language listening are perhaps the least well described and analysed in the currently available literature on language assessment” (Taylor & Geranpayeh, 2013a, p. 326). The research in this thesis adds to our knowledge of cognition in listening assessment by investigating in detail test takers' response processes in relation to single and double play listening tasks in a complex and sophisticated research design involving a total of 322 participants.

The findings of the two studies clearly show that the choice between single and double play is not simply a matter of what test developers deem more practical or what test takers prefer, but has fundamental consequences for the construct that is measured. It was shown that, in addition to changes to the psychometric properties of the listening tests, a broader type of construct validity was impacted by the single play convention. In single play, test takers rely more on lower-order cognitive processes – i.e. on

decoding individual words, clauses, and sentences – and the arguably more construct-relevant higher-order cognitive processes of meaning building in context and meaning construction at discourse level are underrepresented, a finding which is congruent with Field (2015). As Buck points out, in real-life listening “[c]ontext is usually very important. The most obvious context is what the speaker said earlier, as each section of text becomes the context for interpreting later sections” (Buck, 2018, p. XIII). In order to meet these real-life demands, double play is more effective than single play.

Similarly, my results show that in single play test takers use a smaller variety and a smaller relative amount of construct-relevant listening strategies. This appears to be because they need to spend a considerable amount of cognitive resources on finding the answers to the test questions and consequently rely to a large extent on test-taking strategies. The proportion of test-strategic behaviour in relation to all observed metacognitive processing in the single play condition exceeded 45 percent in the stage before listening to the text, 60 percent in the while-listening stage, and 55 percent in the stage after listening to the text. Test takers also clearly struggled with the multimodal demands of the listening tasks, as they had to simultaneously listen to the text, read the questions, think about the answers, and write down the answers, all of this without getting a second chance of understanding the text. This in turn made them feel more stressed, under pressure, and panicked, which further impacted their understanding, a finding which is congruent with research by Brunfaut and Révész (2015), Elkhafaifi (2005), Hembree (1988), Kim (2000), and Winke and Lim (2014).

These effects pose a serious threat to construct validity, however all of them can be alleviated by playing the listening text a second time. I showed that although participants still relied to a large extent on lower-order cognitive processes in double play, the proportion of higher-order cognitive processing increased markedly in double play compared to single play. Particularly during the second play of the double play condition students listened more globally and tried to integrate what they had understood into the context of the speech situation as a whole. In addition, participants used a greater variety and a greater relative amount of construct-relevant listening strategies, again mostly during the second play of double play, although the use of two listening strategies was already higher in the first play of double play compared to single play. As participants knew from the outset that they would hear the listening text a second time, they were able to assign their cognitive resources in a way that allowed them to focus more on understanding the text than on answering the test questions. In

double play the amount of test-strategic behaviour dropped by 11 percent in the stage before listening to the text, by 11 percent in the while-listening stage of the second play, and by 10 percent in the stage after listening to the text, compared to the single play condition. Getting a second chance at understanding also manifestly reduced candidates' anxiety levels, which in turn appears to have freed up their cognitive resources and helped them gain a fuller understanding of the listening text.

In sum, the research in this thesis provides strong evidence that a double play convention can help mitigate two major threats to the validity of listening assessments: construct-irrelevant variance and construct-underrepresentation (Messick, 1995). As summarised in Table 39, in single play test takers rely to a large extent on test-taking strategies and they are also noticeably anxious, both of which introduce construct-irrelevance into test scores. In addition, the relative lack of higher-order cognitive processing and listening-strategic behaviour negatively affects construct-representation. All of these effects are markedly attenuated by playing the listening text a second time.

Table 39: Threats to construct validity of a single play convention compared to a double play convention

Construct-irrelevant variance	Construct-underrepresentation
<i>In single play, test takers...</i>	<i>In single play, test takers...</i>
<ul style="list-style-type: none"> • rely more on test-management strategies • display more test-wise behaviour • are markedly more anxious 	<ul style="list-style-type: none"> • display fewer higher-order cognitive processes • use a smaller number of listening strategies • display less listening-strategic behaviour
<i>...compared to double play.</i>	<i>...compared to double play.</i>

6.4. Balancing priorities in listening assessment

Assessing L2 listening is a highly complex endeavour as test developers need to account for numerous factors which need to be balanced according to different priorities. Thus, with regards to the question of single play and double play, Taylor and Geranpayeh point out that “[a] convincing case can be made for both approaches, depending upon factors such as test purpose, cognitive demand, task consistency, sampling and practicality, all of which reflect the need to balance competing considerations in test design, construction and delivery” (Taylor & Geranpayeh, 2013b, p. 197). This section

considers the research presented in this thesis in light of these “competing considerations” and discusses each of them in turn.

The first factor test developers need to consider when choosing between a single play and double play convention is the purpose of the test. Test purpose is closely linked to authenticity, that is “the degree of correspondence of the characteristics of a given language test task to the features of a TLU (target language use) task” (Bachman & Palmer, 1996, p. 23). It was argued in the introduction of the thesis that the oft-repeated notion of test takers experiencing only single play in most real-life situations seems outdated, particularly in light of increasing technological advances in many contexts including academic and professional domains. While it used to be the case that students at a University generally only had one chance at understanding a lecture, nowadays live lectures are regularly captured and uploaded to virtual learning environments, and whole academic courses are increasingly offered either fully online or in a hybrid form including online and offline content (Sun & Chen, 2016). This gives students the chance to replay recorded lectures should they mishear or miss important information. Similarly, the rise of digital connectivity in professional domains has changed the way billions of people communicate on a day-to-day basis (Graham, Hjorth, & Lehdonvirta, 2017), with an increasing number of interactions happening in an online environment where conversations can be recorded and replayed (Hubbard, 2017, pp. 94–95). In addition to these general trends, the results of the research presented in this thesis suggest that single play in listening assessment is detrimental to construct-representation and leads to increased levels of construct-irrelevant variance. For these reasons, test providers of general L2 proficiency exams such as TOEIC, TOEFL, IELTS, or Pearson General and Pearson Academic, which are widely used for various high-stakes purposes such as immigration, university admission, or work-related decisions, should consider introducing a double play convention in the listening sections of their exams. Currently, most of these tests play listening texts only once. Similarly, teachers and test providers developing listening tests for classroom exams or for national school leaving examinations may want to rethink their practice if they use single play in their listening tests. The case for repeating listening texts in L2 school exams is particularly strong, as double play is traditional and common practice in L2 language classrooms around the world (Field, 2008, p. 159). Using double play in classroom tests and school leaving exams could also be beneficial in terms of washback, as teachers may then use double play in day-to-day classroom activities as well. This in

turn would help promote the construct-beneficial effects of double play in L2 listening instruction.

However, when it comes to test purpose there may also be contexts in which single play is an important part of the construct. For example, in listening tests for aviation, test providers may want to assess understanding based on only one hearing. Although repetition is an internationally accepted convention in radiotelephony airspace communication (referred to as “readback” or “talkback”) (Kim & Elder, 2015, p. 133), it seems that the high-stakes nature of air-traffic controlling warrants the use of single play to ensure that pilots and air-traffic controllers have the ability to immediately understand critical information in emergency situations. Similarly, language tests for health professionals may need to assess listening in a single play convention, as the real-world situations of many health-care workers demand understanding based on only one hearing. For example, as vividly illustrated by Macqueen, Pill, and Knoch (2016), international medical graduates are regularly confronted with the “aural nightmare” of ward rounds, where they need to listen to senior doctors’ conversations with patients in a noisy hospital environment, while simultaneously performing numerous other tasks such as registering information on medical instruments and taking notes of everything they see and hear (Macqueen et al., 2016, p. 281). However, the findings of the current study demonstrate that in contexts where single play is an important part of the construct test providers should use tasks which are less prone to construct-irrelevant factors such as the use of test-wiseness strategies and which may provoke anxiety. The findings presented in this thesis indicate that test takers use higher amounts of test-wiseness strategies in MC than NF tasks. In addition, they were markedly more anxious in MC tasks, which in turn appears to have impacted their understanding. Thus, my results show that single play should be treated carefully as task choice in that condition may be more prone to method effects. A good example of a single play listening task which abides by these principles is Part A of the current Occupational English Test (OET), where test takers have to complete notes during a consultation instead of answering MC questions.

Another consideration when deciding between single and double play is cognitive demand. Ideally, listening tests should only assess cognitive processes and listening strategies which also play a role in real-life listening. As discussed above, I showed that test takers use a larger amount of construct-relevant higher-order cognitive processes and a greater variety and larger amount of listening strategies in double play compared

to single play. Thus, from a purely cognitive perspective, the research in this thesis has shown that double play is superior to single play, which is congruent with findings by Field (2015).

Apart from test purpose and cognitive demand, a third priority of listening test developers is task consistency. Although consistency alone does not tell us about the meaning of test scores, the more reliable test scores are, the more trust we can have in them (Chapelle, 2012). I showed in Study 1 that the overall reliability of test tasks is enhanced by a double play condition. This effect was slightly stronger for MC tasks than NF tasks. In addition, the two task formats were also more comparable with regards to task difficulty in the double play versus the single play condition, suggesting that task format impacts task difficulty less in double play than single play.

Finally, yet another factor that needs to be balanced against competing priorities when deciding whether to use single or double play in listening assessment is task sampling. Task sampling is closely linked to practicality, and a common argument for single play is that test providers can include more tasks in their assessment if the listening texts are played only once. If all listening texts are played twice listening tests would take too much time, so whenever double play is used only a limited number of tasks can be included which in turn negatively affects construct representation, so the argument goes. However, the research presented in this thesis illustrates that a double play convention in fact *enhances* construct representation. I showed that in double play candidates used a greater amount of meaning-building processes and a greater variety and a greater amount of listening strategies compared to single play, and candidates were also less reliant on test-taking strategies and markedly less anxious. If test providers decide to administer listening tests in single play so that they can include more tasks in their assessments, they should be aware that the test results may to a considerable extent be clouded by candidates' reliance on and use of construct-irrelevant test-taking strategies and by their high anxiety levels. Test developers should also acknowledge that in a single play convention, with tasks including multiple items on one coherent listening text, they are mostly testing decoding at word, clause, and sentence level, and that test takers' use of listening strategies is limited compared to double play. In sum, by including more tasks in single play test providers are not capturing more of the construct, but will just capture the same part of the construct repeatedly. Double play broadens the construct in ways that simply adding another task in a single play condition will inevitably miss.

In conclusion, in terms of Taylor and Geranpayeh's "competing priorities" in listening assessment (2013b, p. 197), it seems that double play should be the default condition unless test purpose shows a clear mandate for single play only. The current research provides strong arguments against a single play convention in general L2 proficiency exams. It was demonstrated that double play is beneficial in terms of cognitive demand, task consistency, as well as task sampling.

7. Conclusion

In this final chapter of the thesis the findings will be summarised in terms of the theoretical, methodological, and practical contribution of the thesis. The chapter will also outline the limitations of the research and will conclude with suggestions for further studies.

7.1. Theoretical contribution

The research in this thesis has revealed a number of important and novel theoretical insights not only into the convention of repeating the audio in listening assessment and its implications for the underlying construct, but also into test takers' response processes in listening assessment more generally. By investigating in detail the impact of single and double play on four major groups of response processes – test takers' cognitive processes, listening strategies, test-taking strategies, and anxiety – the two studies in the thesis agree that double play is superior to single play in terms of construct validity as conceptualised by Messick (1989, 1995). By playing the listening text a second time, test developers can mitigate two major threats to the validity of listening tests: construct-irrelevant variance and construct underrepresentation.

With regards to construct-irrelevant variance, a persistent challenge for language test developers are candidates' use of and reliance on test-taking strategies as well as their anxiety (Winke & Lim, 2014). Both of these pose a threat to construct validity, as test scores obtained through test-wiseness or affected by candidates' anxiety are not an accurate reflection of the underlying trait that tests try to measure (Golchi, 2012). It was shown in this thesis that the impact of test-taking strategies and anxiety on listening test scores are significantly smaller in double play compared to single play, particularly in the second play of the double play condition. These advantageous effects were the same across the different tasks and task types (multiple-choice and open format) as well as the various stages of task completion. By hearing the listening text a second time, candidates were able to focus less on answering the test questions, which made them feel more relaxed and less anxious. This in turn was beneficial for their listening comprehension, as high anxiety in single play may have negatively impacted students' understanding.

In terms of construct representation, listening test developers generally try to include as many tasks as feasible within a certain time frame in their listening assessments in order to tap into as much of the construct as possible. For this reason, many international listening test providers utilise single play, which takes less time to administer than double play. By including more tasks, test developers also hope to increase the reliability of their measurements (Fortune, 2004; Green, 2017; Jones, 2011). However, it was shown in this thesis that construct representation and reliability are in fact enhanced in double play compared to single play. While candidates relied to a large extent on lower-order cognitive processing in single play, in the double play condition they displayed a larger amount of construct-relevant higher-order cognitive processes, which are associated with building meaning in context (Buck, 2018; Field, 2013). In terms of metacognitive processing, candidates used a greater variety and a larger relative amount of listening strategies in double play versus single play. These effects were again more prevalent in the second play of the double play condition, however the use of certain listening strategies was already higher in the first play of double play compared to single play. Thus, overall, the listening construct is represented more fully when candidates can hear the recording a second time. It was also shown that test reliability as measured by Cronbach's Alpha is enhanced in double play versus single play for both multiple-choice as well as open format tasks, albeit the effect was slightly larger for multiple-choice tasks.

On a more general level, the analysis of the verbal recall data revealed the incredibly complex interactions between candidates' listening processes, reading processes, test-taking processes, and emotive reactions while completing L2 listening assessment tasks. I showed that L2 listening tests do not only assess test takers' listening comprehension, but also their ability to focus on multiple modalities at the same time, as well as their capacity to deal with feelings of anxiety and stress. It was demonstrated that when the audio is played a second time, L2 listening assessments are much closer to what they are supposed to be: tests of speech comprehension rather than tests of multitasking in a stressful situation.

7.2. Methodological contribution

Similar to previous studies (e.g. Harding, 2011; Rukthong, 2015), the research presented in this thesis has shown the importance of a mixed methods design for

investigating test takers' response processes in L2 listening assessment. A common limitation associated with questionnaire research is that research questions can only be investigated superficially (Dörnyei & Taguchi, 2009, p. 7). Similarly, verbal report data alone is often of limited usefulness as results usually cannot be generalised due to small sample sizes. However, by combining the two methods they complement each other and thereby not only give researchers more confidence in the results, but also offer deeper insights into candidates' response processes which would inevitably be missed when employing each method independently. In the current research, the use of questionnaires to study test takers' strategic behaviour and anxiety levels allowed for the results to be generalised onto a larger population, while the analyses of the verbal recalls confirmed and substantiated the questionnaire findings and offered detailed and personal insights into test takers' response processes from the test takers' perspective. In addition, the analysis of candidates' answer changes by using two different coloured pens in test administration revealed important findings.

In terms of verbal recall procedure, the use of eye-tracking in combination with stimulated recall has proven to be a powerful tool for tapping into test takers' thoughts, as this approach mitigates two common criticisms of verbal report data. First, reactivity effects are minimised because candidates are not interrupted in their test-taking process, but can complete an entire listening task as they would in a normal exam situation before recalling their thoughts. Second, the concrete and graphic reminder of test takers' response processes by replaying a video of their eye-traces overlaid with the audio mitigates memory effects often associated with retrospective recalls, as test takers not only see their eye-movements, but also which answers they chose and at what time in relation to the listening text. In the current study, 15 out of 16 test takers explicitly mentioned that seeing their eye-traces helped them remember their thought processes, more so than only being reminded about their answers would have done.

7.3. Practical contribution

The findings of the research in this thesis are of practical relevance for listening test providers. As outlined in the introductory chapter, current practices of test developers with regards to single and double play in L2 listening assessment vary considerably. While double play is standard practice in classroom assessment and many national school leaving exams, a number of widely used international high-stakes listening tests

utilise single play only. This is concerning in light of the studies presented in this thesis, as it was shown that in single play different parts of the construct are captured compared to double play. In addition, construct-irrelevant factors such as test-taking strategies and anxiety are impacted markedly by the number of times the listening text is played. The current research provides strong arguments for double play to be the default condition in L2 listening tests, as it is beneficial in terms of construct representation, construct irrelevance, cognitive demand, task consistency, as well as task sampling, provided the listening tasks are developed according to state-of-the-art guidelines. Thus, unless test purpose shows a clear mandate for single play only, test developers should strongly consider assessing L2 listening in a double play condition.

Apart from test developers, language teachers could also benefit from the research in this thesis. Although double play is common in teaching L2 listening (Field, 2008, p. 159), language teachers often feel the need to also train students in understanding listening texts in single play, either because they believe that single play is more authentic in terms of real-life listening (Vandergrift & Goh, 2012, pp. 4–5), or because they want to prepare students for exams featuring single play. However, as pointed out by Vandergrift and Goh, “[t]he downside of this practice is that learners are constantly trying to understand what they hear but never get a chance to step back and learn how to deal with the listening input” (Vandergrift & Goh, 2012, p. 5). As demonstrated in this thesis, when the recording is played only once students mostly focus on understanding individual words and sentences and do not pay much attention to the context and overall meaning of the message, as they are too preoccupied with simultaneously answering the test questions.

7.4. Limitations

Despite the carefully planned research design and administration, some limitations of the findings need to be acknowledged. One limitation concerns the unique features and existing policies of the Austrian research context, particularly with regards to the tasks used. All of the tasks in the research were originally developed for a double play convention. Although certain parameters such as instructions and time to check answers were adjusted in the single play condition, the results may not be generalizable to listening tasks developed for single play only. Still, the findings are congruent with Field (2015), who used IELTS listening tasks (developed for single play only) across

single and double play, which indicates that the effects may be similar regardless of which condition the tasks are developed for.

A second limitation in relation to the Austrian research context is the fact that the participants in the two studies were all used to a double play convention, as double play is standard practice in the Austrian school system and the Matura exam. Although double play is also common in many educational settings around the globe (Field, 2008, p. 159; Hubbard, 2017), students who are trained in responding to single play listening tasks may display different response processes.

Another limitation concerns the fact that participants did not complete the tasks in an actual high-stakes exam situation. Although conditions of an exam situation were simulated, students may have displayed different response processes compared to a real test because the test results were of no consequence to them (A. D. Cohen, 2006, p. 313). However, it seems reasonable to hypothesise that in an actual exam situation differences between single play and double play in relation to construct-irrelevant factors, such as anxiety and reliance on test-wiseness strategies, may have been even more pronounced.

Finally, one limitation has to do with collecting data on a stationary eye-tracker, which restricts participants' head movements. In Study 2, 4 out of 16 participants commented that this distracted them slightly, which may have resulted in somewhat different response processes. An alternative might be to use eye-tracking glasses, however their accuracy in terms of eye-movement readings is limited compared to stationary eye-trackers which operate at higher sampling frequencies.

7.5. Suggestions for further research

This thesis has identified a number of areas where further research could potentially help us gain an even fuller understanding not only on the question of single and double play in listening assessment, but also of candidates' response processes in listening assessment more generally. First, an analysis of eye-tracking metrics in combination with verbal recall data may reveal novel insights, particularly with regards to candidates' attention to reading while completing listening assessment tasks. I originally planned to investigate this, however an initial exploration of the data suggests that the link between gaze position and attention may not be as straightforward for listening tasks as is assumed in some studies (e.g. Aryadoust, 2019), but may be

somewhat fraught. During the eye-tracking experiments I noticed that participants would sometimes look at a certain word or part of a sentence for an unusually long time, particularly during the second play of the double play condition. When I asked them about this in the stimulated recalls, they often stated that they did not actually pay attention to what they were looking at, but were only focussing on understanding the listening text. As they did not want to look off screen or close their eyes, they just looked at a random word or part of a sentence, without actually paying attention to it. Thus, it would have been impossible to disentangle candidate's reading processes from their listening processes through eye-tracking data alone, so this line of research was not pursued. Still, further research employing eye-tracking in combination with other process tracing methods such as stimulated recalls would help clarify the link between candidates' gaze position and attention when completing listening assessment tasks.

Another relevant area for further research is the role self-paced listening in L2 listening assessment. In an increasing number of everyday listening situations (e.g. while listening to podcasts, audio books, recorded online conversations, online university lectures, YouTube videos, etc.) listeners are in full control of the recording. In these situations, listeners can not only hear a recording a second time, but they can also re-listen to specific passages, as well as pause, rewind, and fast-forward. In this sense, as argued in Section 6.4, double play in listening assessment seems closer to real life-listening than single play, however even double play is restricted as test takers do not have control over the recording. Future studies may thus want to investigate how self-paced listening could be implemented in listening tests, with a view to cater for the increasing role of technology in real-life listening tasks.

A third important area which would benefit from further research on the role of double play is the assessment of students with specific learning difficulties (SpLDs). In many language tests, students with SpLDs are given extra time to complete the test tasks. For example, in the current version of the IELTS listening test, students with SpLDs may be offered a special version of the test, whereby a supervisor stops the recording to allow more time to read the questions and write down answers. Studies have shown that allowing generous amounts of time for task completion is important for test takers both with and without SpLDs (Cahan, Nirel, & Alkoby, 2016; Kormos & Ratajczak, 2019). However, given the findings of the research in this thesis, students with SpLDs would likely also benefit from double play in listening assessment, perhaps

even more so than students without SpLDs. Future research could thus explore the role of double (or multiple) play in test accommodations.

Finally, the findings could also be extended to other research domains where listening is central to the process of data collection, such as studies in which listeners make judgements of accentedness or comprehensibility (e.g. Isaacs & Trofimovich, 2012; Saito, Trofimovich, & Isaacs, 2017), studies where listeners perform orthographic transcription to measure intelligibility (e.g. Derwing & Munro, 1997; Kang, Thomson, & Moran, 2018), and other studies in the field of SLA research which seek to operationalise listening comprehension (e.g. Krüger, 2018; Levak & Son, 2017). Given the results of my study, it is reasonable to hypothesise that playing stimuli once or twice in research settings is likely to influence not only the type of cognitive processing listeners engage in, but also their strategic behaviour. Such changes have implications for the validity of response processes whether listeners are playing the role of judges, transcribers, or comprehenders. At the very least, the findings of this study demonstrate that providing information on how many times stimuli were played should be an essential reporting requirement in any L2 research which includes listening activities. In the longer term, the effects of single versus double play could be explored across all such contexts to investigate the impact of single, double or multiple play.

7.6. Concluding remarks

Although investigating the effects of single versus double play may seem, at face value, like a niche topic even within the specialised field of language testing, the study presented in this thesis demonstrates that exploring a very specific practical question can open up wide ranging and important insights into the nature of a construct, in this case listening comprehension. The results of this study are useful not only for testing and assessment, but for understanding the way in which L2 listeners process and comprehend speech more generally, and the role of tasks in mediating those processes. This thesis is ultimately a demonstration of how an applied linguistics problem can lead to theory-building which has broad implications for the field. The thesis also demonstrates that continuing to question and explore conventions and orthodoxy can be fruitful in developing clearer understandings of the constructs we work with.

Bibliography

- Afflerbach, P., Pearson, P. D., & Paris, S. G. (2008). Skills and strategies: Their differences, their relationships and why it matters. In K. Mokhtari & R. Sheorey (Eds.), *Reading strategies of first- and second-language learners - See how they read* (pp. 11–24). Norwood, MA: Christopher-Gordon Publishers, Inc.
- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (2000). *Statistics with confidence* (2nd ed.). BMJ Books.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & N. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anderson, J. R. (1995). *Cognitive psychology and its implications* (4th ed.). New York: Freeman.
- Aryadoust, V. (2019). Dynamics of item reading and answer changing in two hearings in a computerized while-listening performance test: an eye-tracking study. *Computer Assisted Language Learning*, 1–28.
<https://doi.org/10.1080/09588221.2019.1574267>
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F. (2004). *Statistical analysis for language assessment*. Cambridge: Cambridge University Press.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Badger, R., & Yan, X. (2012). The use of tactics and strategies by Chinese students in the listening component of IELTS. In L. Taylor & C. J. Weir (Eds.), *IELTS collected papers 2: Research in reading and listening assessment* (pp. 454–486). Cambridge: Cambridge University Press.
- Bartlett, F. C. (1932). *Remembering: A Study in experimental and social psychology*. Cambridge: Cambridge University Press.
- Berne, J. E. (1995). How does varying pre-listening activities affect second language listening comprehension? *Hispania*, 78(2), 316–329.
- Bouroughs, R. (2003). The change process at the paper level. Paper 4: Listening. In C. J. Weir & M. Milanovic (Eds.), *Continuity and innovation. Revising the Cambridge Proficiency in English examination 1913–2002. Studies in language testing 15*. (pp. 315–366). Cambridge: Cambridge University Press.
- Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York: Routledge.

- Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, 19(4), 369–394.
<https://doi.org/10.1191/0265532202lt236oa>
- Brown, J. D. (2001). *Using surveys in language programs*. Cambridge, UK: Cambridge University Press.
- Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes while completing reading tasks: A mixed-method eye-tracking and stimulated recall study*. ARAGs Research Reports Online. London: The British Council.
- Brunfaut, T., & Révész, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly*, 49(1), 141–168.
<https://doi.org/10.1002/tesq.168>
- Buck, G. (1990). *The testing of second language listening comprehension*. Unpublished PhD thesis, University of Lancaster.
- Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing*, 8(1), 67–91. <https://doi.org/10.1177/026553229100800105>
- Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.
- Buck, G. (2018). Preface. In G. J. Ockey & E. Wagner (Eds.), *Assessing L2 listening - Moving towards authenticity* (pp. XI–XVI). Amsterdam / Philadelphia: John Benjamins Publishing Company.
- Cahan, S., Nirel, R., & Alkoby, M. (2016). The extra-examination time granting policy: A reconceptualization. *Journal of Psychoeducational Assessment*, 34(5), 461–472. <https://doi.org/10.1177/0734282915616537>
- Campbell, I. (2007). Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine*, 26, 3661–3675.
- Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, 27(2), 270–295.
- Chang, A. C.-S., & Read, J. (2006). The effects of listening support on the listening performance of EFL learners. *TESOL Quarterly*, 40(2), 375–397. Retrieved from <http://onlinelibrary.wiley.com/doi/10.2307/40264527/abstract>
- Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254–272.
- Chapelle, C. A. (2012). Reliability in language assessment. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Blackwell Publishing Ltd.
<https://doi.org/doi:10.1002/9781405198431.wbeal1003>
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). Building a validity argument for the Test of English as a Foreign Language™. New York: Routledge.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13. <https://doi.org/10.1111/j.1745-3992.2009.00165.x>
- Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In Antony John Kunnan (Ed.), *The companion to language assessment* (pp. 1079–1097). New Jersey: John Wiley and Sons.
<https://doi.org/10.1002/9781118411360.wbcla110>

- Chen, C., Lee, S.-Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, 6, 170–175.
- Cheng, L., & Sun, Y. (2015). Interpreting the impact of the Ontario Secondary School Literacy Test on second language students within an argument-based validation framework. *Language Assessment Quarterly*, 12(1), 50–66.
- Chiang, C. S., & Dunkel, P. (1992). The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL Quarterly*, 26(2), 345–374.
- Cohen, A. D. (1987). Using verbal reports in research on language learning. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research* (pp. 82–95). Philadelphia: Multilingual Matters.
- Cohen, A. D. (1996). Verbal Reports as a source of insights into second language learner strategies. *Applied Language Learning*, 7, 11–27.
- Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, 3(4), 307–331.
<https://doi.org/10.1080/15434300701333129>
- Cohen, A. D. (2011). *Strategies in learning and using a second language* (2nd ed.). Harlow: Pearson Education Limited.
- Cohen, A. D., & Upton, T. A. (2007). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL®. *Language Testing*, 24(2), 209–250. <https://doi.org/10.1177/0265532207076364>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advances in mixed methods research designs. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 209–240). Thousand Oaks, CA: Sage Publications.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Cutler, A., & Clifton, C. (1999). Comprehending spoken language: A blueprint of the listener. In *The Neurocognition of language* (pp. 123–166). Oxford: Oxford University Press.
- Davies, A., & Elder, C. (2005). Validity and validation in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 795–813). Mahwah, NJ: Lawrence Erlbaum Associates.
<https://doi.org/10.1017/s0261444802241828>
- de Winter, J. C. F., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, 39(4), 695–710.
<https://doi.org/10.1080/02664763.2011.610445>

- Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19(1), 1–16.
- Doe, C. C. M., & Fox, J. (2011). Exploring the testing process: Three test takers' observed and reported strategy use over time and testing contexts. *The Canadian Modern Language Review*, 67(1), 29–54. <https://doi.org/10.1353/cml.2010.0034>
- Dörnyei, Z., & Taguchi, T. (2009). *Questionnaires in second language research: Construction, administration, and processing* (2nd ed.). New York: Routledge.
- EALTA. (2006). EALTA guidelines for good practice in language testing and assessment. Retrieved from <http://www.ealta.eu.org/documents/archive/guidelines/English.pdf>
- Eberharter, K., & Frötscher, D. (2012). Quality control in marking open-ended listening and reading test items: Issues and practice. In D. Tsagari, S. Papadima-Sophocleous, & S. Ioannou-Georgiou (Eds.), *International experiences in language testing and assessment*. Frankfurt: Peter Lang.
- Eckes, T. (2015). *Introduction to Many Facet Rasch Measurement* (2nd ed.). Frankfurt am Main: Peter Lang.
- Elkhafaifi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *The Modern Language Journal*, 89(2), 206–220. <https://doi.org/10.1111/j.1540-4781.2005.00275.x>
- Enright, M., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater scoring. *Language Testing*, 27(3), 317–334.
- Ericsson, K. A., & Simon, H. A. (1987). Verbal reports on thinking. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research* (pp. 24–53). Philadelphia: Multilingual Matters.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, Massachusetts: MIT Press.
- Faerch, C., & Kasper, G. (1986). The role of comprehension in second language learning. *Applied Linguistics*, 7(3), 257–274. <https://doi.org/10.1111/j.1467-1770.1976.tb00277.x>
- Field, J. (2008). *Listening in the language classroom*. Cambridge: Cambridge University Press.
- Field, J. (2009). Two bites of the cherry: The effect of replay on the listener. *Paper Presented at BAAL Annual Meeting, Newcastle*.
- Field, J. (2012). The cognitive validity of the lecture-based question in the IELTS listening paper. In L. Taylor & C. J. Weir (Eds.), *IELTS collected papers 2: Research in reading and listening assessment* (pp. 391–453). Cambridge: Cambridge University Press.
- Field, J. (2013). Cognitive validity. In L. Taylor & A. Geranpayeh (Eds.), *Examining listening* (pp. 77–151). Cambridge: Cambridge University Press.
- Field, J. (2015). *The effects of single and double play upon test outcomes and cognitive processing*. ARAGs Research Reports Online. London: The British Council.

- Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 231–235). Hillsdale, N.J.: Lawrence Erlbaum.
- Fortune, A. J. (2004). *Testing listening comprehension in a foreign language – Does the number of times a text is heard affect performance ?* Unpublished MA dissertation: University of Lancaster.
- Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing*, 29(3), 345–369. <https://doi.org/10.1177/0265532211424479>
- Gass, S., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Gass, S., & Mackey, A. (2007). *Data elicitation for second and foreign language research*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 17–30). Englewood Cliffs, NJ: Prentice Hall.
- Goh, C. C. M. (1998). How learners with different listening abilities use comprehension strategies and tactics. *Language Teaching Research*, 2, 124–147.
- Goh, C. C. M. (2002). Exploring listening comprehension tactics and their interaction patterns. *System*, 30(2), 185–206. [https://doi.org/10.1016/S0346-251X\(02\)00004-0](https://doi.org/10.1016/S0346-251X(02)00004-0)
- Golchi, M. M. (2012). Listening anxiety and its relationship with listening strategy use and listening comprehension among Iranian IELTS learners. *International Journal of English Linguistics*, 2(4), 115–128. <https://doi.org/10.5539/ijel.v2n4p115>
- Graham, M., Hjorth, I., & Lehdonvirta, V. (2017). Digital labour and development: Impacts of global digital labour platforms and the gig economy on worker livelihoods. *Transfer: European Review of Labour and Research*, 23(2), 135–162. <https://doi.org/10.1177/1024258916687250>
- Green, A. (1998). *Verbal protocol analysis in language testing research*. Cambridge: Cambridge University Press.
- Green, R. (2013). *Statistical analyses for language testers*. New York: Palgrave Macmillan.
- Green, R. (2017). *Designing listening tests: A practical approach*. London: Palgrave Macmillan. <https://doi.org/10.1057/978-1-349-68771-8>
- Gwet, K. L. (2014). *Handbook of inter-rater reliability - The definitive guide to measuring the extent of agreement among raters* (4th ed.). Gaithersburg: Advanced Analytics, LLC.
- Hambelton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care*, 38, 60–65.
- Harding, L. (2011). *Accent and listening assessment. Language Testing and Evaluation Volume 21*. Frankfurt am Main: Peter Lang.
- Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House.

- Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, 58(1), 47. <https://doi.org/10.2307/1170348>
- Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Cambridge, MA: Newbury House.
- Henning, G. (1991). *A study of the effects of variation of short-term memory load, reading response length, and processing hierarchy on TOEFL listening comprehension item performance (Educational Testing Service, RR 90-18)*. Princeton, NJ: Educational Testing Service.
- Holmquist, K., Nyström, N., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (Eds.). (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.
- Holzknacht, F., Eberharter, K., Kremmel, B., McCray, G., Zehentner, M., Konrad, E., & Spöttl, C. (2017). *Looking into listening: Using eye-tracking to establish the cognitive validity of the Aptis Listening Test*. ARAGs Research Reports Online. London: The British Council.
- Horwitz, E. K. (2010). Foreign and second language anxiety. *Language Teaching*, 43(02), 154. <https://doi.org/10.1017/S026144480999036X>
- Hubbard, P. (2017). Technologies for teaching and learning L2 listening. In C. A. Chapelle & S. Sauro (Eds.), *The handbook of technology and second language teaching and learning* (pp. 93–106). Oxford: Wiley-Blackwell. <https://doi.org/10.1002/9781118914069.ch7>
- Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: setting the stage. In A. M. Hubley & B. D. Zumbo (Eds.), *Understanding and investigating response processes in validation research* (pp. 1–12). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-56129-5_1
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.
- Imura, H. (2007). The listening process: Effects of question types and repetition. *Language Education and Technology*, 44, 75–85.
- Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, 34(3), 475–505. <https://doi.org/10.1017/S0272263112000150>
- Jang, E. E., Wagner, M., & Park, G. (2014). Mixed methods research in language testing and assessment. *Annual Review of Applied Linguistics*, 34. <https://doi.org/10.1017/S0267190514000063>
- Jones, G. (2011). *Research Summary: Once or twice? A critical review of current literature on the question how many times the audio recording should be played in listening comprehension testing items*. Pearson Education Limited. Retrieved from <http://pearsonpte.com/research/Documents/ListenOnceOrTwice.pdf>
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.
- Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319–342.

- Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, 29(1), 3–17. <https://doi.org/10.1177/0265532211417210>
- Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5–17.
- Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, 68(1), 115–146.
- Kaplan, R. M., & Saccuzo, D. P. (1997). *Psychological testing: Principles, applications and issues*. Pacific Grove: Brooks Cole Pub.
- Kasper, G. (1998). Analyzing verbal protocols. *TESOL Quarterly*, 32(2), 356–362.
- Kim, H., & Elder, C. (2015). Interrogating the construct of aviation English: Feedback from test takers in Korea. *Language Testing*, 32(2), 129–149. <https://doi.org/10.1177/0265532214544394>
- Kim, H. J. (2000). *Foreign language listening anxiety: A study of Korean students learning English*. Unpublished PhD thesis: University of Texas.
- Kimura, H. (2008). Foreign language listening anxiety: Its dimensionality and group differences. *JALT Journal*, 30(2). Retrieved from http://jalt-publications.org/files/pdf/jalt_journal/2008b_jj.pdf#page=27
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.
- Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, 35(4), 477–499. <https://doi.org/10.1177/0265532217710049>
- Kormos, J., & Ratajczak, M. (2019). *Time-extension and the second language reading performance of children with different first language literacy profiles*. ARAGs Research Reports Online. London: The British Council.
- Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2005). The measurement of attitudes. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 21–76). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Krüger, M. (2018). Second language acquisition effects of a primary physical education intervention: A pilot study with young refugees. *PLoS ONE*, 13(9), e0203664.
- Lado, R. (1961). *Language testing*. London: Longman.
- Landis, R. J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Levak, N., & Son, J.-B. (2017). Facilitating second language learners' listening comprehension with Second Life and Skype. *ReCALL*, 29(2), 200–218.
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Levelt, W. J. M. (1993). Language use in normal speakers and its disorders. In H. Grimm, J. C. Marshall, & C.-W. Wallesch (Eds.), *Linguistic disorders and pathologies* (pp. 1–15). Berlin: De Gruyter.
- Levelt, W. J. M. (1995). The ability to speak: From intentions to spoken words. *European Review*, 3, 13–23.

- Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago: MESA Press.
- Linacre, J. M. (2019). Dummy facets for interactions. Retrieved June 17, 2019, from <https://www.winsteps.com/facetman/dummy.htm>
- Long, D. R. (1990). What you don't know can't help you - An exploratory study of background knowledge and second language listening comprehension. *Studies in Second Language Acquisition*, 78(1), 65–80.
- Lumley, T., & Brown, A. (2005). Research methods in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 833–855). Mahwah, NJ: Lawrence Erlbaum Associates.
- Lund, R. J. (1991). A comparison of second language listening and reading comprehension. *The Modern Language Journal*, 75(2), 196–204. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/j.1540-4781.1991.tb05350.x/abstract>
- MacIntyre, P. D., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language Learning*, 44(2), 283–305.
- Mackey, A., & Gass, S. (2005). *Second language research: Methodology and design*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Macqueen, S., Pill, J., & Knoch, U. (2016). Language test as boundary object: Perspectives from test users in the healthcare domain. *Language Testing*, 33(2), 271–288. <https://doi.org/10.1177/0265532215607401>
- Magno, C. (2009). Demonstrating the difference between Classical Test Theory and Item Response Theory using derived test data. *The International Journal of Educational and ...*, 1(1), 1–11. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1426043
- Markham, P., & Latham, M. (1987). The influence of religion-specific background knowledge on the listening comprehension of adult second-language student. *Language Learning*, 37(2), 157–170.
- McCray, G., Alderson, J. C., & Brunfaut, T. (2012). *Validity in reading comprehension items: Triangulation of eye-tracking and stimulated recall data*. Paper presented at the EALTA conference: University of Innsbruck, Austria.
- McCray, G., & Brunfaut, T. (2016). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing*. <https://doi.org/10.1177/0265532216677105>
- McNamara, T. (1996). *Measuring second language performance*. London and New York: Longman.
- McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment*. Oxford: Oxford University Press.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
- Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749.

- O'Malley, J. M., & Chamot, A. . (1990). *Learning strategies in second language acquisition*. Cambridge: Cambridge University Press.
- O'Sullivan, B., & Weir, C. J. (2011). Language testing and validation. In B. O'Sullivan (Ed.), *Language testing: Theory and practice* (pp. 13–32). Oxford: Palgrave.
- Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, 24(4), 517–537. <https://doi.org/10.1177/0265532207080771>
- Ockey, G. J., & Wagner, E. (Eds.). (2018). *Assessing L2 listening - Moving towards authenticity*. Amsterdam / Philadelphia: John Benjamins Publishing Company.
- Oller, J. (1979). *Language tests at school*. London: Longman.
- Osborne, J. W., & Costello, A. B. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation*, 10(7), 1–9. <https://doi.org/10.1.1.110.9154>
- Oxford, R. (1990). *Language learning strategies: What every teacher should know*. New York: Newbury House.
- Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, 26(1), 136–144. <https://doi.org/10.7334/psicothema2013.259>
- Pallant, J. (2007). *SPSS survival manual* (3rd ed.). Milton Keynes, UK, USA: Open University Press.
- Purpura, J. E. (2013). Cognition and language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1452–1476). New Jersey: John Wiley and Sons. <https://doi.org/10.1002/9781118411360.wbcla150>
- Quarfoot, D., & Levine, R. A. (2016). How robust are multirater interrater reliability indices to changes in frequency distribution? *The American Statistician*, 70(4), 373–384.
- Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, 35(1), 31–65. <https://doi.org/10.1017/S0272263112000678>
- Richardson, J. T. E. (2011). The analysis of 2 x 2 contingency tables - Yet again. *Statistics in Medicine*, 30, 890.
- Robin, R. (2007). Learner-based listening and technological authenticity. *Language Learning & Technology*, 11(1), 109–115.
- Rost, M. (2011). *Teaching and researching listening* (2nd ed.). Harlow: Pearson Education Limited.
- Rubin, J. (1981). Study of cognitive processes in second language learning. *Applied Linguistics*, 11(2), 117–131. <https://doi.org/10.1093/applin/2.2.117>
- Ruhm, R., Leitner-Jones, C., Kulmhofer, A., Kiefer, T., Mlakar, H., & Itzlinger-Bruneforth, U. (2016). Playing the recording once or twice: Effects on listening test performances. *International Journal of Listening*, 30(1–2), 67–83. <https://doi.org/10.1080/10904018.2015.1104252>
- Rukthong, A. (2015). *Investigating the listening construct underlying listening-to-summarize tasks*. Unpublished PhD thesis: Lancaster University, UK.

- Russo, J., Johnson, E., & Stephens, D. (1989). The validity of verbal protocols. *Memory and Cognition*, 17(6), 759–769. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/2811673>
- Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, 38(4), 439–462. <https://doi.org/10.1093/applin/amv047>
- Sakai, H. (2009). Effect of repetition of exposure and proficiency level in L2 listening tests. *TESOL Quarterly*, 43(2), 360–372. Retrieved from <http://onlinelibrary.wiley.com/doi/10.2307/3586886/abstract>
- Salverda, A. P., Brown, M., & Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual world studies. *Acta Psychologica*, 137(2), 172–180. <https://doi.org/10.1016/j.actpsy.2010.09.010>
- Sanchez, M. E. (1992). Effects of questionnaire design on the quality of survey data. *Public Opinion Quarterly*, 56, 216–217.
- Sarig, G. (1989). Testing meaning construction: Can we do it fairly? *Language Testing*, 6(1), 77–94.
- Sasaki, M. (2013). Introspective methods. In A. J. Kunnan (Ed.), *The companion to language assessment*. New Jersey: John Wiley and Sons. <https://doi.org/10.1002/9781118411360.wbcla076>
- Schmidt-Rinehart, B. C. (1994). The effects of topic familiarity on second language listening comprehension. *The Modern Language Journal*, 78(2), 179–189. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/j.1540-4781.1994.tb02030.x/abstract>
- Schmitt, N. (2002). Do reactions to tests produce changes in the construct measured? *Multivariate Behavioral Research*, 37(1), 105–126.
- Schoonjans, F. (2018). Comparison of proportions calculator. Retrieved July 10, 2018, from https://www.medcalc.org/calc/comparison_of_proportions.php
- Sherman, J. (1997). The effect of question preview in listening comprehension tests. *Language Testing*, 14(2), 185–213. <https://doi.org/10.1177/026553229701400204>
- Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, 84(2), 127–190. <https://doi.org/10.1037/0033-295X.84.2.127>
- Spöttl, C., Eberharter, K., Holzknicht, F., Kremmel, B., & Zehentner, M. (2018). Delivering reform in a high stakes context: From content-based assessment to communicative and competence-based assessment. In G. Sigott (Ed.), *Language testing in Austria: Taking stock (Sprachtesten in Österreich: Eine Bestandsaufnahme)* (pp. 219–240). Frankfurt: Peter Lang.
- Sun, A., & Chen, X. (2016). Online education and its effective practice: A research review. *Journal of Information Technology Education: Research*, 15(September 2015), 157–190. <https://doi.org/10.28945/3502>

- Taylor, L., & Geranpayeh, A. (2013a). Conclusions and recommendations. In L. Taylor & A. Geranpayeh (Eds.), *Examining listening* (pp. 322–341). Cambridge: Cambridge University Press.
- Taylor, L., & Geranpayeh, A. (Eds.). (2013b). *Examining listening*. Cambridge: Cambridge University Press.
- Teng, H. C. (1998). Effects of cultural schemata and visual cues on Chinese students EFL listening comprehension. *Papers from the Eleventh Conference on English Teaching and Learning in the Republic of China*, 550–553.
- Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.
- Turner, C. E. (2013). Mixed methods research. In Antony John Kunnan (Ed.), *The companion to language assessment* (pp. 1162–1180). New Jersey: John Wiley and Sons. <https://doi.org/10.4324/9780203777176>
- Vandergrift, L. (1997). The comprehension strategies of second language (French) listeners: A descriptive study. *Foreign Language Annals*, 30, 387–409.
- Vandergrift, L. (2003). Orchestrating strategy use: Toward a model of the skilled second language listener. *Language Learning*. *Language Learning*, 53, 463–496.
- Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, 40(3), 191. <https://doi.org/10.1017/S0261444807004338>
- Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. New York: Routledge. <https://doi.org/10.4324/9780203843376>
- Vogely, A. J. (1998). Listening comprehension anxiety: students' reported sources and solutions. *Foreign Language Annals*, 31, 67–80. <https://doi.org/http://dx.doi.org/10.1111/j.1944-9720.1998.tb01333.x>
- Vogt, W. (2007). *Quantitative research methods for professionals*. Boston, MA: Pearson.
- Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, 5(3), 218–243. <https://doi.org/10.1080/15434300802213015>
- Wagner, E. (2012). Surveys. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell Publishing Ltd. <https://doi.org/10.1093/oxfordhb/9780199215362.013.28>
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.
- Wenden, A. (1987). Metacognition: An expanded view of the cognitive abilities of L2 learners. *Language Learning*, 37, 573–594.
- Winke, P., & Lim, H. (2014). The effects of testwiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation. *IELTS Research Reports Online Series*, (3), 1–30.
- 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(1), 21–44. <https://doi.org/10.1177/026553229801500102>

- Xi, X. (2008). Methods of test validation. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education, Volume 7: Language testing and assessment* (2nd ed., pp. 177–196). New York: Springer Reference.
- Xie, Q. (2011). Is test taker perception of assessment related to construct validity? *International Journal of Testing*, 11(4), 324–348.
<https://doi.org/10.1080/15305058.2011.589018>
- Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, 32(2), 199–225.
- Young, M. Y. C. (1997). A serial ordering of listening comprehension strategies used by advanced ESL learners in Hong Kong. *Asian Journal of English Language Teaching*, 7, 35–53.
- Zaiontz, C. (2019). Real statistics resource pack. Real Statistics Using Excel.
 Retrieved from <http://www.real-statistics.com/free-download/real-statistics-resource-pack/>

Appendix

1. Ethical consent documents

Dear parents,

Your daughter/your son is invited to take part in a research study. Please take time to read the following information. If you do not wish your daughter / your son to take part in this study, please get in touch with me. If you do not get in touch we assume that you consent to your child taking part in the study.

What is the purpose of this study? Why has my daughter/my son been invited?

I am carrying out this study as part of my Doctoral studies in the Department of Linguistics and English Language at Lancaster University. The aim of the study is to find out what students are thinking while they are solving an English listening test.

What does the study entail?

In the study the participants will complete four English listening tasks and answer a number of short questionnaires. The study will take place during one English lesson.

What are the possible benefits from taking part?

The participants will get feedback on their performance on the listening tasks. The tasks used in the study were developed by experts in test development.

What are the possible disadvantages and risks of taking part?

There are no disadvantages or risks to taking part.

What will happen if I decide not to take part or if I don't want to carry on with the study?

If you decide not to take part in this study, this will not affect your child's studies and the way they are assessed on their course. You are free to withdraw from the study at any time and you do not have to give a reason. If you withdraw while the study takes place or until 2 months after it finishes, I will not use any of the information that you provided. If you withdraw later, I will use the information you shared with me for my study.

Will my taking part in this project be kept confidential?

All the information collected about you during the course of the research will be kept strictly confidential. Any identifying information, such as names and

personal characteristics, will be anonymised in the PhD thesis or any other publications of this research. The data I will collect will be kept securely. Any paper-based data will be kept in a locked cupboard. Electronic data will be stored on a password protected computer and files containing personal data will be encrypted.

What will happen to the results of the research study?

The results of the study will be used for academic purposes only. This will include my PhD thesis and other publications, for example journal articles. I am also planning to present the results of my study at academic conferences.

What if there is a problem?

If you have any queries or if you are unhappy with anything that happens concerning your child's participation in the study, please contact myself or my supervisor.

Further information and contact details

Franz Holzknacht
f.holzknacht1@lancaster.ac.uk
0041 7660 49640

Supervisor:
Luke Harding
l.harding@lancaster.ac.uk
0044 1524 593034

Thank you for considering your child's participation in this project.

Best wishes,
Franz Holzknacht



Consent Form

Project title: Investigating the thought processes of listening test takers

1. I have had explained to me the purposes of the project and what will be required of me, and any questions have been answered to my satisfaction. I agree to the arrangements described in the information sheet in so far as they relate to my participation.
2. I understand that my participation is entirely voluntary and that I have the right to withdraw from the project any time, but no longer than 2 months after its completion. If I withdraw after this period, the information I have provided will be used for the project.
3. I understand that all data collected will be anonymised and that my identity will not be revealed at any point.
4. I have received a copy of this consent form and of the accompanying information sheet.

Name:

Signed:

Date:



2. Questionnaire 1

*Please indicate your agreement to the following statements. Make sure to relate your answers only to **the two tasks you just completed** and where **you heard the recording once**.**

	1	2	3	4	
	Disagree	Partly disagree	Partly agree	Agree	Don't know
1. I read the questions/answer options before listening.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. I predicted my own answer after listening and then looked at the options. <i>[only for MC tasks]</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. I made a guess based on vocabulary used in the questions (and options).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. I listened for the words that appeared in the questions (and options).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. I only listened for relevant information to answer the questions.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. I filled in the answer sheet anyway, though I wasn't not sure.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. Before taking the test, I felt confident and relaxed.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. During the test, I found myself thinking of the consequences of failing.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. During the test, I got so nervous that I forgot facts I really know.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. After taking the test, I felt I could have done better than I actually did.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11. When I first got my copy of the test, it took me a while to calm down to the point where I could begin to think straight.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12. While I took the test, my nervousness caused me to make careless errors.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13. While taking the test, I found myself wondering whether the other students were doing better than I was.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14. I concentrated hard on what the speaker was saying.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15. I guessed the meaning of unknown words, using tone of voice as a clue.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16. While listening, I made up a story line, or adopted a clever perspective.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17. I made a mental or written summary of language and information presented in the listening tasks.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
18. I translated what I heard into my mother tongue.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
19. While listening, I monitored my understanding of the listening passage discourse structure (e.g., compare/contrast, definition).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
20. I got upset when I was not sure whether I understood what I was hearing in English.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
21. I often understood the words but still couldn't quite understand what the speaker was saying.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
22. I got so confused I couldn't remember what I'd heard.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
23. I felt intimidated while listening to the tasks.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
24. I enjoyed listening to the tasks.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
25. I felt confident while listening to the tasks.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Instructions were changed accordingly for the double play condition.*

3. Questionnaire 2

1. Gender		female	<input type="checkbox"/>	male	<input type="checkbox"/>
2. Age		15	<input type="checkbox"/>	16	<input type="checkbox"/>
		17	<input type="checkbox"/>	18	<input type="checkbox"/>
		19	<input type="checkbox"/>		
3. Did you grow up bilingually?		yes	<input type="checkbox"/>	no	<input type="checkbox"/>
4. Mother tongue(s) (multiple answers possible)					
German	<input type="checkbox"/>	English	<input type="checkbox"/>	French	<input type="checkbox"/>
Italian	<input type="checkbox"/>	Turkish	<input type="checkbox"/>	Slovenian	<input type="checkbox"/>
Serbian	<input type="checkbox"/>	Hungarian	<input type="checkbox"/>	other language	<input type="checkbox"/>
5. How familiar were you with the topics of the different tasks?					
		1 not at all familiar		2 rather not familiar	
				3 rather familiar	
					4 very familiar
5.1.	Michael Apted	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>
5.2.	Recycling plastic bottles	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>
5.3.	Swan upping	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>
5.4.	Lego master model builder	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>
6. How difficult did you find the tasks?					
		1 very difficult		2 rather difficult	
				3 rather not difficult	
					4 not difficult
6.1.	Michael Apted	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>
6.2.	Recycling plastic bottles	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>
6.3.	Swan upping	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>
6.4.	Lego master model builder	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>
7. How familiar were you with the test methods used in the test?					
		1 not at all familiar		2 rather not familiar	
				3 rather familiar	
					4 very familiar
7.1.	Multiple Choice (e.g. Michael Apted)	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>
7.2.	Note form (e.g. Swan upping)	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>
8. How well were you able to show your listening competency in the four tasks?					
		1 not well at all		2 rather not well	
				3 rather well	
					4 very well
8.1.	Michael Apted	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>
8.2.	Recycling plastic bottles	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>
8.3.	Swan upping	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>
8.4.	Lego master model builder	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>
9. Which do you prefer?		single play	<input type="checkbox"/>	double play	<input type="checkbox"/>
9.1. Why?					

The layout was changed slightly fit the parameters of the thesis.

4. Study 1: instructions for the test administration

VORGANGSWEISE FÜR DIE TESTUNG

Erstmal vielen Dank für Ihre/Deine Bereitschaft, diese Testung für mich durchzuführen! Dieses Dokument beschreibt den Ablauf der Testung.

Nach Erhalt des Pakets

1. Überprüfen Sie bitte den Inhalt des Testpakets anhand der beigelegten Checkliste (Blatt 2).
2. Vergewissern Sie sich, dass alle Testhefte enthalten und nummeriert sind.
3. Überprüfen Sie bitte, ob die Audiodateien bzw. die CD korrekt abgespielt werden.

Am Tag vor der Testung

1. Teilen Sie das Informationsblatt (Blatt 3) in der Stunde vor der Testung aus und geben Sie den SchülerInnen 5 Minuten Zeit, es zu lesen (in der Stunde der Testung ist dafür leider nicht genug Zeit). Beantworten Sie etwaige Fragen.
2. Machen Sie sich mit dem Ablauf der Testung vertraut, indem Sie dieses Dokument aufmerksam durchlesen.
3. Der Text in den Kästchen muss den KandidatInnen möglichst wörtlich vorgelesen werden, um eine Standardisierung der Testung zu gewährleisten.

Allgemeine Informationen

Die Arbeitszeit für die vier Hörverstehenaufgaben, die Fragebögen im Testheft und den Feedbackfragebogen am Schluss beträgt ca. 45 Minuten. Es wird ohne Pause durchgearbeitet.

Am Tag der Testung, vor Testbeginn

1. Überprüfen Sie bitte, ob die Audiodateien vom jeweiligen Computer (bzw. die CD vom jeweiligen CD-Player) abgespielt werden können. Überprüfen Sie bitte außerdem, ob die Lautsprecher den Testungsraum ausreichend beschallen. Beachten Sie, dass gerade kleine und leistungsschwache Geräte bei hoher Lautstärke den Klang verzerren, wodurch die Verständlichkeit der Aufnahmen gefährdet wird. Organisieren Sie bitte gegebenenfalls ein adäquates Ersatzgerät.
2. Lüften Sie den Raum.
3. Ordnen Sie die Tische so an, dass zwischen den KandidatInnen möglichst viel Platz bleibt, um die Möglichkeit des Abschreibens zu minimieren. Falls vorhanden, verwenden Sie bitte Trennwände.
4. Löschen Sie gegebenenfalls die Tafel.
5. Kleben Sie ein nummeriertes Etikett auf jeden Tisch/Platz.
6. Legen Sie einen roten Kugelschreiber auf jeden Tisch/Platz.
7. Erstellen Sie einen Sitzplan auf dem beigelegten Blatt 4 mit den Nummern der KandidatInnen:

125	126	127
124	123	122
	PULT	

Nach Eintreffen der KandidatInnen

Schließen Sie die Tür. Lesen Sie den KandidatInnen vor:

- Räumt bitte die Tische frei.
- Ihr benötigt zwei verschiedenfarbige Stifte, blau bzw. schwarz und rot. Bitte verwendet euren eigenen blauen bzw. schwarzen Stift. Ein roter Kugelschreiber liegt auf eurem Platz. Diesen könnt ihr behalten.
- Diese Testung ist anonym, jedem Schüler / jeder Schülerin wird eine Nummer, die ihr auf den Etiketten seht, zugeordnet.

Sobald die KandidatInnen ihre Plätze eingenommen haben, überprüfen Sie, dass sich auf den Tischen nur die beiden Stifte und das Etikett befinden.

Die Testung

Lesen Sie den SchülerInnen die einleitenden Informationen vor:

- Bearbeitet diesen Test wie eine normale Schularbeit.
- Notiert euch nach der Testung eure KandidatInnennummer. Anhand dieser Nummer könnt ihr einige Wochen nach der Testung eure Ergebnisse erfahren. Diese werden der Lehrperson zugeschickt.

Lesen Sie den SchülerInnen die Anweisungen zu Ablauf und Durchführung der Testung vor:

- Der Hörverstehentest beinhaltet 4 Aufgaben.
- Zwei der Aufgaben werdet ihr zweimal hören und die anderen zwei Aufgaben nur einmal.
- Nach jeweils zwei Aufgaben werdet ihr einen Fragebogen ausfüllen.
- Am Ende werdet ihr einen Feedback-Fragebogen ausfüllen.
- Verwendet zu Beginn des Tests euren blauen bzw. schwarzen Stift.
- Verwendet den roten Stift nur dann, wenn ihr von der Sprecherin dazu aufgefordert werdet.
- Ihr werdet nur bei jenen Aufgaben, die zweimal abgespielt werden, dazu aufgefordert, den roten Stift zu verwenden, und zwar erst bevor ihr die Aufgabe zum zweiten Mal hört.
- Ihr bekommt nach diesen Aufgaben erneut eine Anweisung, wenn ihr wieder den blauen bzw. schwarzen Stift verwenden sollt.

Teilen Sie die Testhefte aus. **Vergewissern Sie sich, dass die Nummern der Testhefte mit den Kandidatennummern (Etiketten) übereinstimmen.**

- Ich teile nun die Prüfungsunterlagen für den Hörverstehentest aus.
- Unterschreibt als erstes die Einverständniserklärung ganz oben auf der Titelseite. Danach könnt ihr euch die Aufgaben kurz anschauen.
- Wenn es noch Fragen gibt, stellt sie jetzt. Während des Tests dürfen keine Fragen mehr gestellt werden.

Erlauben Sie eine Minute Zeit, die Aufgabenstellungen durchzulesen und beantworten Sie gegebenenfalls Fragen dazu.

- Verwendet zu Beginn den blauen bzw. schwarzen Stift.
- Der Test beginnt jetzt.

1. Starten Sie die erste Audiodatei. Die erste Audiodatei (bzw. der erste Titel auf der CD) enthält Aufgabe 1 und 2. Die zweite Audiodatei (der zweite Titel auf der CD) enthält Aufgabe 3 und 4.
2. Nach den ersten beiden Aufgaben (nachdem die Sprecherin sagt: „This is the end of task 2...“):
 - a. drücken Sie auf „Pause“, falls Sie einen CD-Player verwenden.
 - b. teilen Sie den SchülerInnen mit, dass sie jetzt den ersten Fragebogen ausfüllen sollen. Geben Sie dafür ca. 5 Minuten Zeit.
3. Starten Sie anschließend die zweite Audiodatei (bzw. drücken Sie erneut auf „Play“), um die nächsten beiden Aufgaben abzuspielen.
4. Nach der letzten Aufgabe (nachdem die Sprecherin sagt: „This is the end of the listening test.“), teilen Sie den SchülerInnen mit, dass sie jetzt den zweiten Fragebogen ausfüllen sollen. Geben Sie dafür wieder ca. 5 Minuten Zeit.
5. Füllen Sie während der Testung den Testungsbericht (Blatt 5) aus. Notieren Sie relevante Informationen wie Probleme, Störungen, unangebrachtes Verhalten, unvorhergesehene Vorkommnisse usw. inklusive der Nummern der betreffenden KandidatInnen. Bitte achten Sie während der Testung auch darauf, dass die SchülerInnen die roten Stifte nur dann verwenden, wenn Sie dazu aufgefordert werden (ausschließlich beim zweiten Durchlauf jener Aufgaben, welche zweimal abgespielt werden).

Nach Ablauf der Zeit lesen Sie folgende Zeilen vor:

- Die Testung ist jetzt vorbei.
- Hört bitte auf zu schreiben.
- Ich werde jetzt die Unterlagen einsammeln.
- Bis alle Unterlagen eingesammelt sind, ist das Sprechen nicht erlaubt.

Feedbackfragebögen

Nachdem die Testhefte eingesammelt wurden, teilen Sie die Feedbackfragebögen aus. **Vergewissern Sie sich, dass die Nummern der Feedbackfragebögen mit den Kandidatennummern (Etiketten) übereinstimmen.**

- Bitte füllt nun den Feedbackfragebogen aus.
- Für das Ausfüllen des Fragebogens werden etwa 5 Minuten veranschlagt.

Nach ca. 5 Minuten sammeln Sie die Fragebögen ein. Zählen Sie sämtliche Unterlagen. Erst danach können die SchülerInnen entlassen werden.

- Herzlichen Dank für die Teilnahme und ein erfolgreiches verbleibendes Schuljahr.

Was tun, wenn...

- ein/e KandidatIn zu spät kommt?
 - Der/die betreffende KandidatIn kann an der Testung eingeschränkt teilnehmen, solange gewährleistet werden kann, dass die anderen KandidatInnen dadurch nicht gestört werden. Die Verspätung bitte sowohl auf dem Deckblatt des Testheftes als auch im Testungsbericht vermerken.
- ein/e KandidatIn aus gesundheitlichen Gründen den Raum verlassen muss (Übelkeit, Nasenbluten etc.)?
 - Falls der/die betreffende KandidatIn die Testung vollständig abbrechen muss, das Testheft einsammeln und sowohl auf dem Deckblatt des Testheftes als auch im Testungsbericht vermerken.
 - Falls der/die betreffende KandidatIn die Testung fortsetzen kann, die Zeit der Absenz sowohl auf dem Testheft als auch im Testungsbericht vermerken.
- KandidatInnen früher fertig sind?
 - Die KandidatInnen müssen das Testheft umdrehen und ruhig in der Klasse warten.
- KandidatInnen nach Ablauf der Zeit noch nicht fertig sind?
 - Die Testhefte absammeln und die Fragebögen austeilen. Es ist nicht erlaubt, länger Zeit zu geben, weil dadurch die Ergebnisse verfälscht werden.
- die Audiodatei fehlerhaft ist?
 - Versuchen Sie durch Vor- oder Zurückspulen den Fehler zu überbrücken und vermerken Sie genau die Stelle (Minutenangabe) im Testungsbericht.
- Testhefte fehlerhaft sind?
 - Bei gravierenden Fehlern (z.B. fehlende Seiten im Testheft): die betreffenden KandidatInnen sollen die Aufgaben bestmöglich bearbeiten. Bitte verständigen Sie mich umgehend nach der Testung (0660 7611652)
 - Bei leichten Fehlern (z.B. Tippfehler), nur auf dem Testungsbericht vermerken (nicht die Testung unterbrechen oder auf den Fehler aufmerksam machen).

Nach der Testung

1. Bitte vervollständigen und unterschreiben Sie die Checkliste.
2. Vervollständigen Sie den Testungsbericht und geben Sie ihn mit der Checkliste, dem ausgefüllten Sitzplan, der CD und diesem Dokument in die Klarsichthülle.
3. Geben Sie alle Testungsmaterialien in das Paket. Es erleichtert meine Aufgabe sehr, wenn die Unterlagen in folgender Reihenfolge verpackt werden:

- Klarsichthülle mit Checklisten, Sitzplänen, CDs, Testungsberichten und diesem Dokument

Nach Kandidatennummern geordnet:

- Ausgefüllte Feedbackfragebögen Gruppe 1
 - Ausgefüllte Testhefte Gruppe 1
 - Leere Feedbackfragebögen Gruppe 1
 - Leere Testhefte Gruppe 1
 - Ausgefüllte Feedbackfragebögen Gruppe 2
 - Ausgefüllte Testhefte Gruppe 2
 - Leere Feedbackfragebögen Gruppe 2
 - Leere Testhefte Gruppe 2
 - (Weitere Gruppen analog dazu)
4. Bitte retournieren Sie das Paket nach Abschluss aller Ihrer Testungen per Post „unfrei“ (Porto zahlt Empfänger) an: Franz Holz knecht, [REDACTED], [REDACTED]

VIELEN DANK FÜR IHRE/DEINE UNTERSTÜTZUNG!

5. Study 1: seating plan for the test administration

SITZPLAN

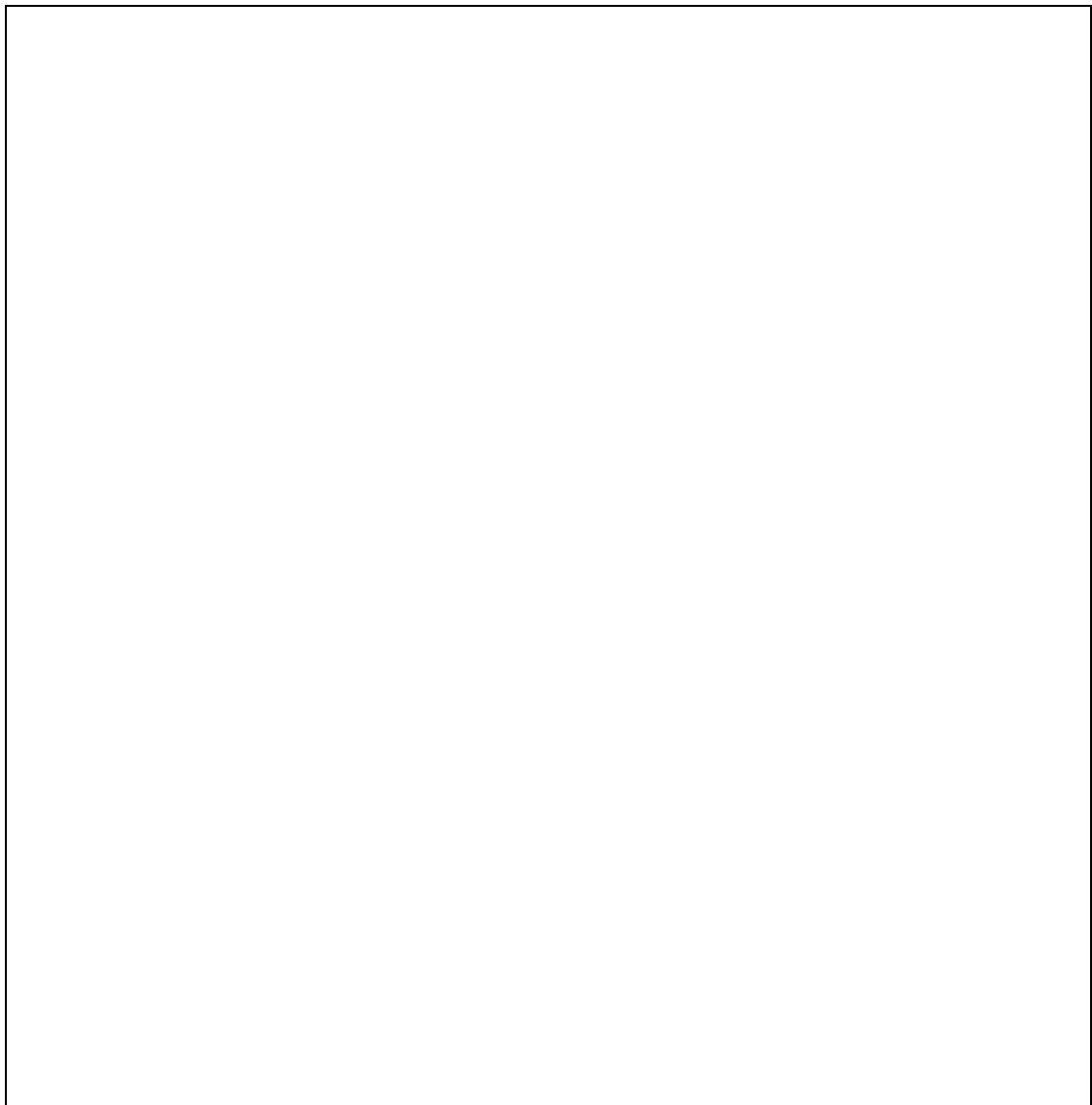
Lehrperson:

Datum:

Ort (Schule, Klasse): _____

Anzahl der KandidatInnen: _____

Bitte zeichnen Sie hier den Sitzplan der Klasse:



6. Study 1: test administration report

TESTUNGSBERICHT

Lehrperson: _____ Datum: _____

Ort (Schule, Klasse): _____ Anzahl der KandidatInnen: _____

1. Gab es Probleme mit den Testheften oder der CD bzw. den Audiodateien, z.B. fehlende Seiten, falsch gedruckt, nicht klar, Brennfehler, Fehler in der Audiodatei usw.? Bitte geben Sie Details an:

2. Waren die Tische angemessen angeordnet, mit genügend Platz für die KandidatInnen zum Schreiben?

3. Wurden Trennwände verwendet? ☐ ja ☐ nein

4. Gab es Störungen, die die Leistung der KandidatInnen hätte beeinflussen können, z.B. Lärm, Unterbrechungen usw.? Bitte geben Sie Details an:

5. Sind KandidatInnen zu spät gekommen? Wenn ja, welche (Nummern)?

6. Hatten die KandidatInnen Fragen, z.B. in Bezug auf Anweisungen, Testmethoden, bestimmte Aufgaben usw.? Bitte geben Sie Details an:

7. Verhielten sich KandidatInnen unangebracht? Wenn ja, welche (Nummern)?

8. Gab es Probleme mit den Fragebögen, z.B. fehlende Seiten, falsch gedruckt, nicht klar usw.? Bitte geben Sie Details an:

9. Haben die KandidatInnen die roten Stifte nur dann verwendet, wenn sie dazu aufgefordert wurden?

☐ ja ☐ nein

10. Sonstige Vorkommnisse:

Allgemeine Kommentare

7. Study 2: double-coding document (excerpt)

Double-coding

Instructions

Please use the coding scheme to assign a thematic code to each of the data segments below. Type the code into the “code/comments” column.

The data segments are drawn from retrospective recalls, stimulated recalls, and post-hoc interviews with 16 listeners about their listening experience.

If you cannot decide which code to assign, please explain your thoughts in the code/comments box.

N	data segment	code/comments
0	<i>Genau. Da hat er irgendwas mit „cinema“ gesagt und ich so: Was mit „cinema“ genau, also irgendwie so eine Vokabelsache wahrscheinlich oder vielleicht kenne ich das Wort und ich habe es auf jeden Fall nicht verstanden.</i>	<i>lexical search</i>
1	Also ich weiß nicht, es war dann irgendwie schon mit/also ich habe zwar alles gehört, aber ich habe nichts ausgefüllt. Und dann war das irgendwie/ja/„if all people are in fact“ [sic]/ja also ich glaube, das mit dem Astronauten war da noch nicht oder?	
2	Ja es war ja, ich weiß nicht genau, wie er es gesagt hat, aber er hat gesagt, dass er gedacht hat, dass sie auch ziemlich, ich weiß nicht, wie er es gesagt hat, nicht „horrible“ [sic], aber halt nervig war. Und dann irgendwie ist es dann logisch, dass es dann „he changed for the better“ [sic] war	
3	Ja da habe ich mir dann gedacht, ja ok „Blazer“ [sic], das muss dann irgendwo da sein, aber es war dann doch nicht da, denke ich.	
4	Ja. Da irgendwie/ich bin mir immer noch nicht ganz sicher, weil er hat irgendwie von viel später geredet, aber er als er gesagt hat, dass sie in Vorbild ist und dass sie halt geheiratet hat und dass es irgendwie/also dass sie sich halt schon irgendwie gebessert hat.	
5	Also dieses „defined by class“ [sic] habe ich komplett gar nicht verstanden, das hat mich auch verwirrt.	
6	Da habe ich mich jetzt ein bisschen gestresst. Und auch nicht wirklich gewusst, was ich jetzt ankreuzen soll. Und ja, eben bei 11 und 12 habe ich nicht mehr wirklich mitgehört, weil ich halt mit den Gedanken ein bisschen wo anders war.	

8. Study 2: data excerpts

Retrospective recall: P07 MC2 twice retrospective

INT: Wie ist es dir jetzt gegangen bei dem? #00:00:06-0#

P07: Ja war ein bisschen anders, weil man sich halt doch auf den Computer konzentrieren muss. Aber/ #00:00:14-4#

INT: Hat dich das irritiert ein bisschen? Oder nicht? #00:00:15-6#

P07: Ja vor allem, dass man seinen Kopf nicht bewegen kann, aber sonst überhaupt nicht. #00:00:22-9#

INT: Aber das mir dem Kopf hat dich schon ein bisschen gestört? #00:00:24-0#

P07: Ja normalerweise schaut man halt schon in der Klasse herum. #00:00:29-4#

INT: Ok. Und was hast du dir dabei gedacht wie du die Aufgabe gehört hast? #00:00:35-8#

P07: Ja ich habe halt zuerst die Fragen durchgelesen und dann immer geschaut, ob ich etwas Ähnliches finde wie die Frage sagt, so was ähnliches. #00:00:50-7#

INT: Hast du irgendwelche Schwierigkeiten gehabt die Aufnahme zu verstehen? #00:00:53-0#

P07: Nein. Eigentlich nicht. #00:00:56-9#

INT: Das erste Mal auch schon nicht oder generell nicht? #00:01:01-5#

P07: Also sicher merkt man sich nicht gleich alles, aber beim zweiten Mal/ich habe es halt so gemacht, dass wie ich es zum ersten Mal gehört habe ein paar Fragen beantwortet habe und beim zweiten Mal dann den Rest oder die, die noch übrig geblieben sind. #00:01:17-1#

INT: Ok. Und wie hast du zugehört? #00:01:24-4#

P07: Halb auf die Fragen geachtet und halb auf das, was sie geredet haben. #00:01:31-0#

INT: Ok. #00:01:33-1#

P07: So ja. #00:01:34-8#

INT: Hast du alles verstanden, was sie gesagt haben? #00:01:34-8#

P07: Ja, das was/halt das was/die „meaning“ [sic] halt das habe ich verstanden. #00:01:48-4#

Stimulated recall: P02 NF1 once task

INT: Was hast du dir gedacht, wie der Anfang der Aufgabe gekommen ist? #00:06:32-4#

P02: Also da merkt man ja immer, inwiefern das jetzt übernommen worden ist, also ob das jetzt wortwörtlich ist oder ob man etwas ändern muss. Und dann habe ich eben dieses „conservation card“ [sic] und dann habe ich mir gedacht, inwiefern das, was dasteht, ähnlich ist zu dem ist, was gesagt worden ist, weil das auch das ist, ob du dich wirklich konzentrieren musst oder ob du die Sätze eins zu eins übernehmen musst. Normal musst du sie wirklich herausuchen und da habe ich mir schon gedacht, ok das hätte ich jetzt nicht hinbekommen. Und ja. #00:07:02-1#

INT: Und wie hast du hingehört, wie hast du zugehört oder wie hast du versucht, das zu verstehen? #00:07:08-7#

P02: Also jetzt gar nicht so/also ich habe jetzt nicht auf den Inhalt vom Verstehen her geschaut, sondern wirklich nur auf die Wörter. Und ja. Also worum es jetzt wirklich gegangen ist/so genau habe ich da jetzt nicht hingehört. #00:07:25-1#

INT: Also du hast nur versucht die Wörter zu hören. #00:07:26-1#

P02: Ja. #00:07:30-5#

INT: Hast du alles verstanden? #00:07:42-8#

P02: Ja eben. Da war ich jetzt erst bei der ersten Phase und da ist es eben schon um das gegangen und dann habe ich den Zusammenhang nicht ganz verstanden und dann habe ich mir noch einmal das Wort hingeschrieben und mir gedacht, vielleicht kommt noch etwas Ähnliches oder ich merke es mir später und dann habe ich glaube ich eh noch etwas aufgeschrieben. Und aber das habe ich mir dann einmal vermerkt und man hat ja dann noch die 60 Sekunden und dann habe ich mir vorgenommen, dass ich es mir da noch einmal so richtig in Erinnerung rufe und schaue, wie es war. #00:08:06-5#

INT: Ok. Wie hast du da jetzt zugehört? #00:08:39-4#

P02: Also da habe ich als erstes dieses „fishing ...“ [sic] gehört und dann habe ich mir schon gedacht, ob es das ist, aber da war ich mir noch nicht so ganz sicher, und dann ist es weiter gegangen und dann ist es noch einmal gekommen, dass es eben das „...“ [sic] ist. Und dann habe ich mir gedacht, dass es schon reichen wird und habe das hingeschrieben. #00:08:56-1#

INT: Was hast du dir da/wie hast du da zugehört? #00:09:37-8#

P02: Also da habe ich schon zugehört, habe mir aber gedacht, dass irgendwie nichts wirklich passt, weil es ist zwar immer um dieses Prinzip gegangen, Kindern in der Schule etwas zu erklären, aber es ist nicht wirklich ein Grund gekommen, was sie vermeiden wollen. Und dann habe ich mir halt so gedacht, ja entweder habe ich es jetzt komplett verpasst, oder es ist noch gar nicht vorgekommen. Ja dann habe ich mir gedacht, jetzt höre ich einmal weiter zu und dann ist es glaube ich eh gekommen. #00:09:58-7#

INT: Und hast du alles verstanden, was er da jetzt sagt? #00:10:00-3#

P02: Ja. #00:10:00-3#

INT: Ok. #00:10:03-9#

#00:10:12-4#

INT: Also da hast du jetzt etwas verstanden. Und was ist dir da jetzt durch den Kopf gegangen genau, weißt du das noch? #00:10:30-4#

P02: Nein nicht wirklich. #00:10:31-8#

INT: Ok. #00:10:33-6#
#00:10:56-4#

INT: Was ist dir da jetzt durch den Kopf gegangen? Wie hast du da jetzt zugehört? #00:10:59-9#

P02: Ja also da war noch ein Wort vor „uniforms“ [sic], das habe ich nicht ganz verstanden, und dann bin ich noch total auf dem Wort geblieben, deswegen habe ich es erst so spät hingeschrieben. Und dann habe ich gedacht, hoffentlich ist es nicht so relevant, ich schreibe

jetzt einfach einmal nur „uniforms“ [sic] hin, damit ich zumindest einmal abgesichert bin und ja. Das mit den „six teams“ [sic] habe ich irgendwie komplett ausgeschaltet, weil das war für mich gar nicht relevant, und habe mich wirklich auf das, was sie tragen fixiert. Und ja. #00:11:29-3#

INT: Und wie hast du zugehört, hast du jetzt mehr auf den Inhalt gehört? #00:11:34-5#

P02: Ja in dem Fall habe ich schon auf den Inhalt gehört, aber das eine Wort habe ich dann eben nicht verstanden, also akustisch, das war/das habe ich auch jetzt nicht verstanden. Also/ #00:11:43-3#

INT: „appropriate“ [sic] #00:11:45-7#
#00:12:09-2#

INT: Da bist du lange draufgeblieben auf dem Wort? #00:12:12-4#

P02: Ja weil ich nicht gewusst habe, wie ich das jetzt hinschreiben soll, weil ich das Wort in dem Zusammenhang gar nicht gekannt habe. Jetzt habe ich auch nicht gewusst, was das richtige gewesen wäre, und ich hätte nicht einmal gewusst, was ich hätte hinschreiben können. Deswegen ist das dann auch leer geblieben und war dann auch ein bisschen verwirrt und bin erst dann zur nächsten Frage gegangen. #00:12:29-9#

INT: Ok. Und wie hast du da jetzt hingehört, bei der Passage genau? #00:12:32-9#

P02: Also da habe ich eigentlich schon recht genau hingehört, vor allem deswegen, weil der Anfang recht leicht verständlich war einfach. Und dann hat der am Schluss irgendwie so ein bisschen schneller geredet, kommt mir jetzt persönlich vor, ich weiß es nicht, ob das stimmt. Und dann ist es für mich so unter gegangen. Und ja. #00:12:49-5#

INT: Ok. #00:12:52-6#

#00:13:26-7#

INT: Wie hast du da jetzt zugehört, was ist dir da jetzt durch den Kopf gegangen? #00:13:28-7#

P02: Ja also da ist dieses „give one answer“ [sic] und dann habe ich jetzt damit gerechnet, dass jetzt etwas aufgezählt wird, dass es so ist, das sollst du jetzt machen. Und dann ist das eher so, das mit dem, da sind sie vorbei und dann dass sie umkreisen und so es war echt so, glaube ich, so eine Und dann war das irgendwie so, ja ok, was schreibe ich jetzt hin, weil das andere hätte ich in vier Wörtern nicht hinbekommen, und dann habe ich gedacht, dass ich das schreibe, bevor ich gar nichts schreibe, weil ich es irgendwie nicht so ganz gecheckt habe. #00:13:54-9#

INT: Ok. Und hast du alles verstanden, was er da jetzt gesagt hat? #00:13:57-0#

P02: Ja. #00:13:57-0#

INT: Schon? Ok. #00:14:01-5#

#00:14:23-9#

INT: Warst du dir da sicher? #00:14:27-3#

P02: Da habe ich es gar nicht gehört. Ich habe kein einziges Mal irgendetwas dazu gehört. Jetzt habe ich nicht gewusst, ob das mit dem ersten zusammenhängt, ob ich den Übergang verpasst habe, weil ich habe irgendwie „riverbank“ [sic] nicht gehört. #00:14:35-9#

INT: Du hast immer auf „Riverbank“ [sic] gewartet? #00:14:35-9#

P02: Ja. Und das ist nicht gekommen. Wahrscheinlich war es gar nicht da und deswegen wird es nicht/ #00:14:44-1#

INT: Und dann bei der nächsten? #00:14:44-8#

P02: Ja da/da war ich eben noch bei der Frage und dann habe ich komplett verpasst, dass es weiter gegangen ist und dann ist auf einmal „Prince... Castle“ [sic] gekommen und ich so, ok. #00:14:55-3#

INT: Ok. cool. #00:15:00-8#

#00:15:27-7#

INT: Und da jetzt am Schluss, wie hast du da zugehört? #00:15:29-4#

P02: Also ich habe irgendwie schon fast damit gerechnet, dass sie der „Queen“ [sic] gehören. Und das habe ich dann zum Glück nicht hingeschrieben. Und dann habe sie das recht lange ... (das habe ich nicht verstanden) und dann habe ich mir gedacht, ok das war es jetzt mit der „Queen“ [sic] und es kommt gar nicht mehr. Und dann haben sie ja noch gesagt, wem es jetzt wirklich gehört. Und das war dann, finde ich, eh noch recht gut zum Verstehen. #00:15:50-5#

INT: Ok. Super. #00:15:56-9#

Post-hoc interview: P14 final questions

INT: Ok. Super. Jetzt sind wir fast fertig. Wie ist es dir jetzt gegangen bei dem Experiment heute? #00:16:27-3#

P14: Ja. Eigentlich wo ich hergekommen bin, hätte ich nicht gewusst, dass da nur Listening kommt und habe auch dann, als ich das festgestellt habe/ich weiß schon, dass ich da Schwierigkeiten habe. Also eher hinzuhören bei solche Aufnahmen. Und ja. Eigentlich ich habe alles verstanden sinngemäß, Details ein bisschen ja. #00:16:59-0#

INT: Ok. Was magst du lieber einmal Hören oder zwei Mal Hören? #00:17:03-4#

P14: Ahm. Ich muss ganz ehrlich sagen, ich habe auch irgendwie, indem ich anders zugehört habe, habe ich mir mehr gemerkt/also ich habe eben, beim zweiten zwei Mal Zuhören, habe ich mich eben nur mehr auf die Fragen/auf die Fragen eingegangen/also was ich beantwortet habe eigentlich und weil ich die Zeit dafür habe. Ich kann es mir leisten, nicht zu verstehen, was der Sinn davon ist, deswegen höre ich nur hin, was die Antwort sein könnte und falls das/und falls ich nur das höre und den Sinn nicht verstehe, wie ich es da angeführt habe, dann ich beim zweiten Mal genauer hinhören. Also beim nur einmal/also beim zweiten Mal sozusagen Hinhören habe ich gleich eigentlich versucht komplett alles zu verstehen und ist auch irgendwie gegangen. Also ja. Ich würde nicht sagen besser, aber ich habe es mir gleich merken können, was der Sinn von dem ist. Ist halt dann ein bisschen schneller gegangen, aber/ #00:18:04-8#

INT: Und was würdest du jetzt bevorzugen bei einem Test, bei einer Schularbeit? #00:18:06-2#

P14: Ich bevorzuge da eigentlich immer zwei Mal, weil ich da immer so eine Absicherung habe, wenn ich es nicht verstehe. Meistens ist es dann eh so, dass ich das, was ich beim ersten Mal nicht verstehe, beim zweiten Mal genauso wenig verstehe. Also was ich beim ersten Mal beantworte oder nicht beantworte und mir merke, was vielleicht die Antwort sein könnte, ändert sich meistens bis zum zweiten Hinhören eh nicht, aber es ist halt so ein Gefühl der Sicherheit. Also vielleicht würde ich das zwei Mal Hinhören bevorzugen. #00:18:38-1#

INT: Wie war es für dich das Video von deinen Augenbewegungen zu sehen? Hat dich das verwirrt, war es hilfreich oder hat es keinen Unterschied gemacht, dich daran zu erinnern, was dir durch den Kopf gegangen ist? #00:18:48-8#

P14: Ja interessant war zu sehen, wo ich am meisten hängen geblieben bin. Also dass ich zum Beispiel unnötigerweise viel zu lange beim zweiten Mal, also bei der schon beantworteten Frage hängen geblieben bin. Also so interessante Sachen einheitlich. Und habe dann eigentlich schon gewusst, wie es ungefähr aussieht, wo ich hinschaue, nach dem ersten Mal und habe dann auch dementsprechend/ #00:19:19-7#

INT: Angepasst. #00:19:20-2#

P14: Ja. Also es war schon gut zu wissen, was da eigentlich ja/was ich genau anschau. Und interessant eigentlich, weil ich hätte nicht gedacht, dass ich so viele anschau, dass ich so viel hin und her/meistens versuche ich genau ein Wort/also wenn ich zum Beispiel einen Satz nicht verstanden habe, weiß ich, welches Wort ich nicht verstanden habe und dann suche ich eigentlich im ganzen Satz dieses Wort und dann merke ich es gar nicht, aber schaue irgendwie den ganzen Satz durch, obwohl das, was rauskommt, ist nur das einen Wort, was ich nicht verstanden habe. Also war das dann eben ein bisschen verwirrend. #00:19:58-7#

INT: Und hat dir das geholfen, dich daran zu erinnern, was dir durch den Kopf gegangen ist? Die Augenbewegungen zu sehen. #00:20:06-3#

P14: Also eindeutig eigentlich, wenn ich etwas nicht verstanden habe, dann habe ich gewusst, warum ich das nicht verstanden habe und was genau, welches Wort. Also ja, sich zu erinnern, wo ich hinschaue und warum ich hinschaue. Eben das war dann leichter. #00:20:23-9#

INT: Super. Vielen Dank.

9. Study 1: exemplary Facets specifications file

```
Facets = 5
Positive = 2,3,4
Noncentered= 1
; Vertical =
Arrange = mN
Models = ?,?B,?B,?B,?,D
*
Labels=
1, students
All student numbers were listed here
*
2, task types, D ; dummy to avoid subsets
1=MC
2=NF
*
3, tasks, D ; dummy to avoid subsets
1=MC1
2=MC2
3=NF1
4=NF2
*
4, conditions, D ; dummy for bias analysis
1=single play
2=double play
*
5, Items; grouped for subtotals
111,111,0, 1
112,112,0, 1
113,113,0, 1
114,114,0, 1
115,115,0, 1
116,116,0, 1
121,121,0, 2
122,122,0, 2
123,123,0, 2
124,124,0, 2
125,125,0, 2
126,126,0, 2
211,211,0, 3
212,212,0, 3
213,213,0, 3
214,214,0, 3
215,215,0, 3
216,216,0, 3
217,217,0, 3
218,218,0, 3
219,219,0, 3
221,221,0, 4
222,222,0, 4
223,223,0, 4
224,224,0, 4
225,225,0, 4
226,226,0, 4
227,227,0, 4
228,228,0, 4
229,229,0, 4
*
data= The data was inserted here
```

10. Response frequencies for Questionnaire 2

Sub-group 1 (MC double play and NF single play)

q5.1 How familiar were you with the topic in MC1?

		Frequency	Percent	Valid Percent
Valid	not familiar at all	95	62.1	62.1
	not really familiar	41	26.8	26.8
	somewhat familiar	15	9.8	9.8
	very familiar	2	1.3	1.3
	Total	153	100.0	100.0

q5.2 How familiar were you with the topic in MC2?

		Frequency	Percent	Valid Percent
Valid	not familiar at all	28	18.3	18.4
	not really familiar	40	26.1	26.3
	somewhat familiar	60	39.2	39.5
	very familiar	24	15.7	15.8
	Total	152	99.3	100.0
Missing	no answer	1	.7	
Total		153	100.0	

q5.3 How familiar were you with the topic in NF1?

		Frequency	Percent	Valid Percent
Valid	not familiar at all	100	65.4	65.8
	not really familiar	29	19.0	19.1
	somewhat familiar	12	7.8	7.9
	very familiar	11	7.2	7.2
	Total	152	99.3	100.0
Missing	no answer	1	.7	
Total		153	100.0	

q5.4 How familiar were you with the topic in NF2?

		Frequency	Percent	Valid Percent
Valid	not familiar at all	78	51.0	51.3
	not really familiar	42	27.5	27.6
	somewhat familiar	27	17.6	17.8
	very familiar	5	3.3	3.3
	Total	152	99.3	100.0
Missing	no answer	1	.7	
Total		153	100.0	

q6.1 How difficult did you find the listening passage in MC1?

		Frequency	Percent	Valid Percent
Valid	very difficult	11	7.2	7.2
	somewhat difficult	49	32.0	32.2
	not really difficult	75	49.0	49.3
	not difficult at all	17	11.1	11.2
	Total	152	99.3	100.0
Missing	no answer	1	.7	
Total		153	100.0	

q6.2 How difficult did you find the listening passage in MC2?

		Frequency	Percent	Valid Percent
Valid	very difficult	15	9.8	9.8
	somewhat difficult	56	36.6	36.6
	not really difficult	65	42.5	42.5
	not difficult at all	17	11.1	11.1
	Total	153	100.0	100.0

q6.3 How difficult did you find the listening passage in NF1?

		Frequency	Percent	Valid Percent
Valid	very difficult	42	27.5	27.5
	somewhat difficult	68	44.4	44.4
	not really difficult	26	17.0	17.0
	not difficult at all	17	11.1	11.1
	Total	153	100.0	100.0

q6.4 How difficult did you find the listening passage in NF2?

		Frequency	Percent	Valid Percent
Valid	very difficult	66	43.1	43.1
	somewhat difficult	64	41.8	41.8
	not really difficult	22	14.4	14.4
	not difficult at all	1	.7	.7
	Total	153	100.0	100.0

q7.1 How familiar were you with the task type MC?

		Frequency	Percent	Valid Percent
Valid	not really familiar	5	3.3	3.3
	somewhat familiar	36	23.5	23.7
	very familiar	111	72.5	73.0
	Total	152	99.3	100.0
Missing	no answer	1	.7	
Total		153	100.0	

q7.2 How familiar were you with the task type NF?

		Frequency	Percent	Valid Percent
Valid	not familiar at all	6	3.9	3.9
	not really familiar	26	17.0	17.0
	somewhat familiar	56	36.6	36.6
	very familiar	65	42.5	42.5
	Total	153	100.0	100.0

q8.1 How well were you able to show your English listening skills in MC1?

		Frequency	Percent	Valid Percent
Valid	not well at all	9	5.9	6.0
	not really well	41	26.8	27.2
	somewhat well	83	54.2	55.0
	very well	18	11.8	11.9
	Total	151	98.7	100.0
Missing	no answer	2	1.3	
Total		153	100.0	

q8.2 How well were you able to show your English listening skills in MC2?

		Frequency	Percent	Valid Percent
Valid	not well at all	7	4.6	4.6
	not really well	45	29.4	29.8
	somewhat well	84	54.9	55.6
	very well	15	9.8	9.9
	Total	151	98.7	100.0
Missing	no answer	2	1.3	
Total		153	100.0	

q8.3 How well were you able to show your English listening skills in NF1?

		Frequency	Percent	Valid Percent
Valid	not well at all	30	19.6	19.9
	not really well	69	45.1	45.7
	somewhat well	44	28.8	29.1
	very well	8	5.2	5.3
	Total	151	98.7	100.0
Missing	no answer	2	1.3	
Total		153	100.0	

q8.4 How well were you able to show your English listening skills in NF2?

		Frequency	Percent	Valid Percent
Valid	not well at all	48	31.4	31.8
	not really well	71	46.4	47.0
	somewhat well	27	17.6	17.9
	very well	5	3.3	3.3
	Total	151	98.7	100.0
Missing	no answer	2	1.3	
Total		153	100.0	

Sub-group 2 (MC single play and NF double play)

q5.1 How familiar were you with the topic in MC1?

		Frequency	Percent	Valid Percent
Valid	not familiar at all	95	62.1	62.1
	not really familiar	38	24.8	24.8
	somewhat familiar	14	9.2	9.2
	very familiar	6	3.9	3.9
	Total	153	100.0	100.0

q5.2 How familiar were you with the topic in MC2?

		Frequency	Percent	Valid Percent
Valid	not familiar at all	15	9.8	9.8
	not really familiar	45	29.4	29.4
	somewhat familiar	74	48.4	48.4
	very familiar	19	12.4	12.4
	Total	153	100.0	100.0

q5.3 How familiar were you with the topic in NF1?

		Frequency	Percent	Valid Percent
Valid	not familiar at all	92	60.1	60.1
	not really familiar	31	20.3	20.3
	somewhat familiar	25	16.3	16.3
	very familiar	5	3.3	3.3
	Total	153	100.0	100.0

q5.4 How familiar were you with the topic in NF2?

		Frequency	Percent	Valid Percent
Valid	not familiar at all	54	35.3	35.3
	not really familiar	50	32.7	32.7
	somewhat familiar	43	28.1	28.1
	very familiar	6	3.9	3.9
	Total	153	100.0	100.0

q6.1 How difficult did you find the listening passage in MC1?

		Frequency	Percent	Valid Percent
Valid	very difficult	13	8.5	8.5
	somewhat difficult	63	41.2	41.2
	not really difficult	61	39.9	39.9
	not difficult at all	16	10.5	10.5
	Total	153	100.0	100.0

q6.2 How difficult did you find the listening passage in MC2?

		Frequency	Percent	Valid Percent
Valid	very difficult	14	9.2	9.2
	somewhat difficult	60	39.2	39.2
	not really difficult	64	41.8	41.8
	not difficult at all	15	9.8	9.8
	Total	153	100.0	100.0

q6.3 How difficult did you find the listening passage in NF1?

		Frequency	Percent	Valid Percent
Valid	very difficult	10	6.5	6.5
	somewhat difficult	38	24.8	24.8
	not really difficult	66	43.1	43.1
	not difficult at all	39	25.5	25.5
	Total	153	100.0	100.0

q6.4 How difficult did you find the listening passage in NF2?

		Frequency	Percent	Valid Percent
Valid	very difficult	16	10.5	10.5
	somewhat difficult	44	28.8	28.8
	not really difficult	60	39.2	39.2
	not difficult at all	33	21.6	21.6
	Total	153	100.0	100.0

q7.1 How familiar were you with the task type MC?

		Frequency	Percent	Valid Percent
Valid	not really familiar	3	2.0	2.0
	somewhat familiar	21	13.7	13.7
	very familiar	129	84.3	84.3
	Total	153	100.0	100.0

q7.2 How familiar were you with the task type NF?

		Frequency	Percent	Valid Percent
Valid	not familiar at all	1	.7	.7
	not really familiar	4	2.6	2.6
	somewhat familiar	27	17.6	17.6
	very familiar	121	79.1	79.1
	Total	153	100.0	100.0

q8.1 How well were you able to show your English listening skills in MC1?

		Frequency	Percent	Valid Percent
Valid	not well at all	13	8.5	8.5
	not really well	62	40.5	40.5
	somewhat well	62	40.5	40.5
	very well	16	10.5	10.5
	Total	153	100.0	100.0

q8.2 How well were you able to show your English listening skills in MC2?

		Frequency	Percent	Valid Percent
Valid	not well at all	8	5.2	5.2
	not really well	61	39.9	39.9
	somewhat well	71	46.4	46.4
	very well	13	8.5	8.5
	Total	153	100.0	100.0

q8.3 How well were you able to show your English listening skills in NF1?

		Frequency	Percent	Valid Percent
Valid	not well at all	5	3.3	3.3
	not really well	27	17.6	17.6
	somewhat well	85	55.6	55.6
	very well	36	23.5	23.5
	Total	153	100.0	100.0

q8.4 How well were you able to show your English listening skills in NF2?

		Frequency	Percent	Valid Percent
Valid	not well at all	12	7.8	7.8
	not really well	40	26.1	26.1
	somewhat well	68	44.4	44.4
	very well	33	21.6	21.6
	Total	153	100.0	100.0

11. Summary tables of observed response processes per participant

Number of quotations coded as **cognitive processes** for each participant:

	lexical search	parsing	meaning construction	discourse construction	Total
P01	14	18	9	1	42
P02	17	11	7	1	36
P03	8	5	3	1	17
P04	12	12	7	3	34
P05	7	15	11	0	33
P06	12	3	4	0	19
P07	4	3	8	0	15
P08	12	6	8	0	26
P09	7	5	6	0	18
P10	22	12	7	0	41
P11	6	4	6	1	17
P12	2	3	0	0	5
P13	2	4	4	1	11
P14	11	7	5	0	23
P15	7	7	3	1	18
P16	7	1	3	0	11
Total	150	116	91	9	366

Number of quotations coded as **listening strategies** for each participant:

	plann.	foc.at.	monit.	eval.	infer.	elab.	pred.	cont.	transl.	m.em.	Total
P01	6	7	3	1	2	9	3	3	2	2	47
P02	8	1	3	3	1	1	4	1	1	1	33
P03	2	7	8	6	0	1	4	0	0	2	30
P04	2	6	0	2	7	3	2	0	0	5	27
P05	1	5	4	3	4	0	1	1	0	1	20
P06	1	9	2	8	2	1	4	0	0	2	29
P07	1	4	2	4	1	0	2	1	0	0	15
P08	5	5	7	3	1	0	3	1	1	0	26
P09	4	9	7	3	1	0	2	0	0	4	30
P10	7	6	9	8	4	4	3	0	0	6	47
P11	3	2	2	2	1	3	1	0	1	0	15
P12	2	5	1	4	0	1	2	0	0	0	15
P13	1	2	3	4	1	0	3	0	0	2	16
P14	1	5	5	6	0	0	4	1	0	0	31
P15	1	5	7	5	1	2	2	0	0	0	23
P16	5	2	4	4	0	0	2	0	0	0	17
Total	59	89	67	75	26	25	42	8	5	25	421

Number of quotations coded as **test-taking strategies** for each participant:

	test- management	test- wiseness	Total
P01	47	1	48
P02	35	3	38
P03	32	0	32
P04	30	0	30
P05	35	0	35
P06	30	2	32
P07	26	1	27
P08	39	2	41
P09	31	3	34
P10	38	1	39
P11	27	4	31
P12	24	1	25
P13	20	3	23
P14	49	1	50
P15	32	0	32
P16	25	1	26

Total	520	23	543
-------	-----	----	-----

Number of quotations coded as **anxiety** for each participant:

	anxiety
P01	4
P02	1
P03	1
P04	4
P05	2
P06	4
P07	1
P08	4
P09	3
P10	4
P11	2
P12	0
P13	2
P14	0
P15	1
P16	1
Total	34