Comparing Perceived and Actual Task and Text Difficulty in the Assessment of Listening Comprehension

Elizabeth Apostolou
University of Athens, Athens

Abstract

The paper reports on a study which investigated compatibility between candidates' perceptions of task and text difficulty in listening comprehension tests and candidate performance in those tests. To this end, a comparison was made between item analysis data derived from six listening comprehension test papers and candidate responses to task and text difficulty as indicated from the analysis of feedback questionnaires concerning the same tests. The study was conducted in the context of the Greek State Certificate of English Language Proficiency, known as KPG, and the data was derived from the B2 level English exam. Through the comparative analyses, candidates' perceived task and text difficulties in the KPG listening tests were found to correlate to a great extent with the results of item analysis, with reference to the same tests. The only cases of inconsistency pertained to the role of paralinguistic features of the oral texts (i.e., speaker's accent) and cognitive variables (i.e., background knowledge) in test performance. Implications are drawn for test developers and item writers of listening comprehension tests as well as for language teachers.

1 Introduction

This paper draws on a broader research project exploring the effect that specific task and text variables can have on the outcome of the listening comprehension process. Language testers have long held an interest in the factors that affect second language test performance and several empirical studies have demonstrated that test score variation in language tests can be attributed to a number of underlying factors. Bachman (1990) and later Bachman and Palmer (1996) proposed a framework for investigating the factors which can affect candidate performance in language tests. They recognized three central categories of factors, i.e., test method characteristics, language ability and the characteristics of test takers. My research seeks to provide empirical evidence in terms of the first set of factors in Bachman and Palmer's (1996) framework (i.e., test method). This area of study is closely linked to their view that

> *since we cannot totally eliminate the effect of task characteristics, we must learn to understand them and control them so as to ensure that the tests we use will have the qualities we desire and are appropriate for the uses for which they are intended (Bachman & Palmer, 1996, p. 46).*

A review of the relevant literature reveals the existence of a series of studies (e.g., Buck, Tatsuoka, Kostin & Phelps, 1997; Buck & Tatsuoka, 1998; Freedle & Kostin, 1999; Spelberg, de Boer & van de Bos, 2000) that have demonstrated the way that specific aspects of the listening test can be associated with overall listening comprehension difficulty. These studies have mainly drawn their findings by examining test scores through psychometric measurement tools such as item analysis. There is, however, a dearth of research focusing on the investigation of task and text difficulty from candidates' perspective, thus using the two sources of information (i.e., test scores and candidates' perspective) comparatively.

The present study is thus motivated by a lack of empirical research on candidates' perceived task and text difficulties in listening comprehension tests and by the ensuing need to explore the extent to which these perceptions correlate with their actual performance in the tests. To address the purpose of this research, data from post administration item analysis of listening test items has been compared and contrasted with candidates' perceptions of listening task and text difficulty obtained from the analysis of feedback questionnaires, with reference to the same examination periods. The purpose of the study is to shed light on the variables of difficulty influencing candidate performance in listening comprehension tests. Discovering which of the candidates' perceived difficulties affect test scores and performance would be a valuable source of information for item writers who wish to design reliable and valid tests.

In the following section, research relevant to this study is reviewed. This includes a discussion on the complex nature of the listening comprehension process and the effect of task- and text-related factors on listening test difficulty. Section 3 describes the context, the participants, the instruments and the research procedure that was used to collect and analyse the data in this study. In sections 4 and 5, results from the comparisons of item analysis data with questionnaire analysis data are presented and discussed. The final section presents directions for future research and reflects on the strengths and limitations of this study.

2 Background to the Study

Although, up to present, very few studies have attempted to shed light on learners' beliefs about listening, most evidence tends to suggest that learners have negative feelings about listening more than they have about any of the other language skills. Arnold (2000) comments on how listening induces anxiety in learners because of the pressure it places on them to process input rapidly (cited in Graham, 2006, pp. 165-166). Graham (2004) investigated language learning in England and she found that for intermediate (i.e., B2 level) learners, listening was the skill in which they experienced the greatest difficulty.

In many ways it is unsurprising that learners perceive listening as difficult. Researchers agree that listening is a complex, active procedure that requires simultaneous use of knowledge, processing skills and strategies. They argue that it is an inferential process in which the listener must use a wider variety of knowledge sources, linguistic and non-linguistic to interpret rapidly incoming data (Anderson & Lynch, 1988; Buck, 2001; Rost, 1990). Buck (2001) explains that listening comprehension involves discrete elements of language such as phonology, vocabulary and syntax but it goes beyond this because it also involves interpretation. Rost (1990) further argues that listening involves background knowledge and listener-specific variables as meaning is constructed within the listener's background and in relation to the listener's purpose. What is more, the listening input is characterized by such features as speech rate, accent, elision, the placement of stress and intonation, redundancy and hesitation, which are unique to listening and different from one language to another (Buck, 2001).

Graham's research (2006) is one of the few that looked at learners' beliefs about listening providing some useful evidence with regards to a) how learners perceive themselves as listeners and to what they attribute their success or lack of it and b) the strategies they were aware of employing when listening. She concluded that many students tend to hold the belief that they are simply not good at listening. The main difficulties identified are coping with

speed of delivery of texts, making out individual words in a stream of spoken language and making sense of any words that have been identified or understood. Furthermore, most learners attribute their difficulties in listening to their supposed low ability in the skill and to the difficulty of the listening tasks and texts. When it comes to strategy use, learners display little insight into what strategies might be appropriate for listening. In general, they express doubts that the problem of task difficulty can be tackled by appropriate listening strategies.

Other research into the testing of listening has mainly focused on the investigation of specific text and task characteristics that may affect difficulty (Brindley & Slatyer, 2002; Freedle & Fellbaum, 1987; Freedle & Kostin, 1999; Jensen, Hansen, Green & Akey, 1997; Nissan, DeVincenzi & Tang, 1996; Shohamy & Inbar, 1991). These studies have highlighted features of task and text that might influence test takers' performance in listening tests. Freedle and Fellbaum (1987) (cited in Yanagawa & Green, 2008, p. 112) found that item difficulty was influenced by the relationship between the text and the answer options. Test items could be made more difficult by increasing the number of lexical repetitions among the incorrect options and by decreasing the number of lexical repetitions among the correct options. Items could also be made more difficult if more lexical inferences were added to the incorrect response options.

Nissan et al. (1996) investigated listening test items taken from 15 TOEFL tests administered before 1996. The study identified five significant predictor variables affecting the difficulty of dialogue test items. These were inference question, utterance pattern, negatives in the text, speaker's role and infrequent vocabulary. In a follow-up to Nissan et al. (1996), Freedle and Kostin (1999) examined the effect of the multiple-choice listening task-type on test difficulty in the TOEFL examination. Among other things, they found that the two most important determinants of difficulty were the location of the necessary information and the degree of lexical overlap. Therefore, when the necessary information came near the beginning of the text or when it was repeated, the item tended to be easier. Similarly, lexical overlap between the text and the correct option was found to be the best predictor of easy items whereas lexical overlap between the text and the incorrect options was the best predictor of difficult items presumably because test-takers tend to select options which contain words they recognize from the passage.

Moreover, Shohamy and Inbar (1991) looked at the effect of three types of questions a) global questions which required test-takers to synthesize information or draw conclusions, b) local questions which required test-takers to locate details or understand individual words, and c) trivial questions which required test-takers to understand precise but irrelevant details not related to the main topic. They found that the global questions were

harder than the local questions but the trivial questions behaved in an unpredictable manner and "served no meaningful purpose in an evaluation tool" (Shohamy & Inbar, 1991, p. 37). Therefore, this research suggests that questions need to focus on key information in the text, not on irrelevant detail.

Another important study on item difficulty was carried out by Jensen et al. (1997) (cited in Brindley & Slatyer, 2002, p. 387). Jensen et al. investigated the effects of text and item characteristics on item difficulty in an English for Academic Purposes listening test. What they discovered was a significant correlation between type of response and length of text; as the length of the text decreased, items requiring a verbatim response – as opposed to a nonverbatim response – became easier.

Furthermore, Brindley and Slatyer (2002) investigated item difficulty in the context of the Certificates in Spoken and Written English (CSWE) assessment system in Australia. They looked into the effect of three structural components of listening assessment tasks on difficulty, namely, the necessary information, the surrounding text and the stem (i.e., item question). Their study suggested that there is a complex interaction between these different components of the task. As a result, particular combinations of item characteristics appear either to accentuate or attenuate the effect on difficulty. For example, they found that an easy, high frequency, one-word response item may become more difficult by the complex syntax of the stem, the weak match with the cue and a long retention time (Brindley & Slatyer, 2002, p. 387).

The above investigations discuss difficulty in terms of specific text and item characteristics. However, in terms of exploring language test performance, we would ideally want to see whether these or other characteristics are also related to what test takers perceive as difficult in a test. In this way, we would be able to determine what it is that causes difficulty to the candidates. The present study is directed towards the above aim by taking into consideration candidates' perspective on task and text difficulty so as to more fully investigate this phenomenon. The following section turns into the context of this study addressing the data collection and research procedure.

3 Method

3.1 Context and Participants

The data presented in this study was collected from the listening comprehension exam papers administered by the Greek State Certificate of English Language Proficiency,

nationally and internationally known as KPG (an acronym for the Greek title Kratiko Pistopiitiko Glossomathias).

The KPG is a high-stakes exam in Greece and as such it can influence one's future prospects for employment and education. It is especially designed for Greek users of the English language and takes into account the social circumstances for its use. It is the only language examination battery in Greece that aims to fulfil the communicative, social, vocational and educational needs of people living, working and studying in Greece. It was developed by taking into account the *Common European Framework of Reference for Languages* (CEFR) (2001) respecting that this document provides a common basis for the recognition of qualifications in all member states. Responsibility for administering the exam lies with the Greek Ministry of Education in collaboration with Departments of Foreign Language and Literature of the National and Kapodistrian University of Athens and the Aristotle University of Thessaloniki.

The exam in English was first introduced in November 2003 at the B2 level. Since then three more levels have gradually been introduced, namely the C1 level (since April 2005), the B1 level (since May 2007), and the A1/A2 level (in one graded written test since May 2008) on the scale set by the CEFR (2001). For the current study, all data were derived from the B2 level as it has been the level attracting the majority of candidates sitting for the KPG English exam since it was first introduced. The B2 level exam is mainly addressed to learners who are above 15 years of age. Eligible to take these exams are EU citizens and others who are living, studying and/or working in Greece and who have a basic knowledge of Greek.

In terms of the development of the exams, the university English team has produced clear specifications regarding the purpose of the exam, its content and intended audience, number of modules, duration and grading (http://rcel.enl.uoa.gr) (www.kpg.minedu.gov.gr). The English language exam consists of four modules each aiming at the assessment of one of the four language skills: Module 1 is entitled *Reading comprehension and language awareness*, Module 2 is entitled *Writing and written mediation*, Module 3 constitutes the *Listening comprehension* test paper and Module 4 is the *Speaking and spoken mediation* test paper (KPG Common Test Specifications, 2007).

The B2 level listening test paper (Module 3) is a 20-minute test. It consists of three to four activities and a total number of 20 test items. The item format includes multiple-choice questions with three options and short answers. The listening activities in general aim at assessing comprehension of the gist of the text, specific ideas in the whole text and in parts of it, what is directly stated or implied and what certain words or expressions mean in the specific context. The listening stimuli at this level are either authentic texts (recorded live,

from the radio, from CDs, or from the Internet) or simulated situations recorded in a studio. Texts are usually heard twice at normal speed and pace. The pronunciation is generally that of natural speakers of English using a standard variety of the target language and clearly articulated sounds. Text types include narrations, radio/TV programmes, news, advertisements, announcements, interviews and everyday conversations. Before listening to the texts, candidates are always given time to read the relevant test questions (KPG B2 test specifications, 2007).

3.2 Data

Data were made available to me through the Research Centre for English Language Teaching, Testing and Assessment (RCEL). RCEL is a unit of the Faculty of English Studies, University of Athens, which, among other things, is responsible for the development of the KPG exams in English. The KPG exam data analysed for the purposes of the present study were collected using two main methodological tools, namely classical item analysis using Item Response Theory (IRT) and feedback questionnaires especially designed for candidates who have sat for the KPG listening test at the B2 level.

Item analysis: In order to determine the degree of exam difficulty, the KPG test development team carries out systematic analysis of the test items after each administration as a means to ensure test validity and consistency.

Item analysis is conducted through specialized software called ITEMAN. ITEMAN is simple in its operation: the user (usually a statistician who works for the RCEL) enters specific examination data from each administration (i.e., exam level, candidates ID, and candidates' responses to test items) and the programme provides automatically, through complex mathematical procedures, the following information: a) internal consistency or reliability of the exam (utilizing Cronbach Alpha), b) index of difficulty (i.e., a value showing the proportion of candidates answering an item correctly), c) discrimination efficiency (i.e., a value showing how well an item succeeds in distinguishing highly competent from less competent candidates) and d) distractor analysis (i.e., the frequency with which each option of a particular test question is chosen).

In terms of item difficulty, the test development team identified as normal values of difficulty for a test item a range between 0.40 and 0.80. This means that any test item that the item analysis shows to have an index of difficulty above 0.80 or below 0.40 is considered to be too easy or too difficult respectively for the exam level. Thus, further analysis is required to determine what makes the specific test item unacceptable for the exam level.

Candidate feedback questionnaires: Since 2004, candidate feedback questionnaires have randomly been distributed to a number of exam centres all over Greece after administering each examination. Their content refers to the reading, writing and listening comprehension test papers of the KPG English exam. Candidates are requested, among other things, to evaluate the level of difficulty of each of these tests. The questionnaires, which are in the form of Likert scales, are provided in the candidates' first language (i.e. Greek).

The questions focusing on the listening comprehension test paper (see Appendix A) aimed at providing information with regards to candidates' perceptions of task and text difficulty in the tests. In particular, candidates were requested to provide feedback on the difficulty level of certain aspects of the listening task and text. With regards to the aspects of the listening task, candidates were asked to rate the level of difficulty of the rubrics included for each activity as well as decide whether it was the stem of the multiple choice test items or the options provided (i.e., the distractors) that caused greater difficulty for them.

The majority of the questions, however, focused on the evaluation of the exam listening stimuli and of certain text difficulty variables as it was not possible to obtain such information from item analysis. More specifically, candidates were requested to rate the level of difficulty of linguistic factors (i.e., lexical difficulty), paralinguistic factors (i.e., speaker accent and rate of speech), cognitive factors (i.e., topic unfamiliarity and lack of background knowledge)[4] and other affective factors (i.e., lack of topic interest and anxiety).

Another reason why the majority of the questions included in the feedback questionnaires focused on the evaluation of text difficulty variables rather than on the characteristics of the listening tasks is that during the pilot phases, when the questionnaires were still at the stage of design, it proved pointless to ask candidates to provide feedback on specific task-related features (i.e., length of question, information organisation, syntactical organisation, lexical overlap, lexical difficulty and inference) that they could hardly recognize in the items or remember after the end of the test.

In the present section, the two basic methodological tools used in my study, i.e., item analysis research and candidate feedback questionnaires were introduced and fully described. I will now proceed with the actual research procedure which, as it will be shown below, was carried out in three phases.

[4] The identification of background knowledge as a cognitive type variable follows Purpura's (1999) conclusion that linking new information with prior knowledge constitutes a cognitive process-type variable representing the storing or memory processes in human information processing.

3.3 Procedure

In the context of the present study, data were collected from six B2 level KPG listening comprehension tests and analysed. These were derived from the examination periods of May 2006, November 2006, May 2007, November 2007, May 2008 and May 2009.

The research procedure involved three steps. The first step concerned the examination of the data derived from item analysis. More specifically, the items that item analysis showed to have demonstrated unacceptable values of difficulty, that is, either above 0.80 or below 0.40, were selected and separated from the effective ones as they were considered to be too easy or too difficult respectively for the exam level. The two categories of 'problematic' test items (i.e., too easy and too difficult) from each examination period were further examined so that conclusions could be drawn as to what features rendered each item difficult or easy for the specific group of candidates. Part of this analysis also involved examination of the incorrect answers (i.e., distractors) included in a particular 'problematic' test question. Alderson, Clapham and Wall (1995) have investigated the role of the distractors in multiple choice test items and found that a low discrimination index can often be explained by the performance of one or more distractors. In the present study the role of the distractors in the 'problematic' test items was examined with the purpose of finding some interaction between item difficulty and distractor performance.

This investigation was complemented with a systematic examination of the texts from which the tasks originated in an attempt to find the relationship between text variables and item difficulty. The analysis concerned linguistic features of the text and especially lexical appropriacy to exam level, information structure, information density, and paralinguistic features (i.e., accent, speech rate, background noise and number of speakers involved) that can have an impact on the level of difficulty of the relevant test items.

As a second step, the analysis of the candidate feedback questionnaires followed. A total number of 6,000 questionnaires corresponding to the aforementioned examination periods were analysed. The following table presents the exact number of B2 level questionnaires analysed from each of the six examination periods:

TABLE 1: B2 level candidate questionnaires

Examination Period	Number of candidate feedback questionnaires
May 2006	B2 level ► 500 questionnaires
November 2006	B2 level ► 494 questionnaires
May 2007	B2 level ► 1000 questionnaires
November 2007	B2 level ► 750 questionnaires
May 2008	B2 level ► 750 questionnaires
May 2009	B2 level ► 2505 questionnaires

The last step involved the comparative analysis that yielded the results of the present study. Therefore, a comparison was drawn between the test items showing an index of difficulty higher than expected (i.e., dif. index<0.40) based on the results of item analysis and candidates' evaluations of certain task and text difficulty variables showing high percentages of difficulty (i.e., above 50%) as indicated by the analysis of the feedback questionnaires. In other words, with reference to the same listening tests, the most difficult (or easy) test items, as displayed by item analysis, were selected to be compared with specific task and text characteristics that were rated as very difficult (or too easy) by more than 50% of the respondents in the questionnaires. The ultimate purpose was to provide findings in terms of whether candidates' perceptions of listening task and text difficulty were consistent with the data derived from item response analysis.

In the following two sections, I will present and interpret the results derived from the third phase of the research process, i.e., the comparative analysis, while in Section 6 the conclusions along with the implications and possible limitations of the current study will be discussed.

3.4 Presentation of Results

Lack of space makes it impossible to present, describe and comment on the findings derived from the analysis of 6,000 candidate feedback questionnaires corresponding to the six KPG listening tests selected in the context of the current research. As a result, I am going to focus

on the results elicited from the investigation of the KPG listening tests administered in May and November 2006 and May 2008, in which a total number of 1,744 candidate feedback questionnaires was collected and analysed (see Table 1 above). The selection of the particular administrations is attributed mainly to two reasons: a) because, based on item analysis, the particular test papers involve a significant number of 'problematic' test items (i.e., either too easy or too difficult) and b) because the analysis of the relevant questionnaires has also shown candidates to face great difficulty with specific text characteristics (see Appendix B, tables 2a-4b).

Regarding data organisation, two tables are provided for each exam period: one illustrating the 'problematic' listening test items with their corresponding difficulty indices (see Appendix B, tables 2a, 3a and 4a) and the other showing candidates' evaluations of specific text characteristics in terms of level of difficulty (see Appendix B, tables 2b, 3b, 4b). Each pair of tables is designed to be read comparatively and contrastively. For this reason, data are categorised in terms of the listening activities they refer to as well as in terms of the oral texts they are associated with. Additionally, correlations between item analysis data and questionnaire results are highlighted so as to be more easily identified by readers. Going through the questionnaire analysis tables (see Appendix B, tables 2b, 3b and 4b), the reader should bear in mind that the numbers do not equal 100%. This is because some respondents did not answer all the questions provided. The 'no answer' parameter was taken into consideration in the analysis but it has been excluded from the tables of this paper to achieve a more accurate illustration of candidates' responses to each question.

Results from each pair of tables (see Appendix B, tables 2a and 2b, 3a and 3b, 4a and 4b) reveal that the respondents' rating of the difficulty level of the oral texts in terms of lexical difficulty, background knowledge and topic interest correlates with the difficulty values indicated by item analysis with reference to the same texts (see for example Appendix B, tables 2a and 2b). As it is evident, the difficult test items for each oral text demonstrate the same hierarchical order of difficulty as candidates' rating of the oral texts in terms of lexical difficulty: *South France* (53%), followed by the *Movie extract* text (52.6%), followed by the texts *Language Museum* (48.2%) and *Radio News* (43.4%) respectively.

This finding leads to the conclusion that candidates' performance in listening can actually be influenced by the level of difficulty they attach to the oral texts accompanying the test items. In particular, specific text-related factors of difficulty (i.e., lexical difficulty), or cognitive (i.e., lack of background knowledge) and affective attributes (i.e., lack of topic interest) can predispose the candidates negatively towards the test items, thus leading to unsuccessful performance. The obvious question arises as to whether the extent to which these difficulty

variables can have an impact on candidates' performance in listening tests. Further research seems to be needed to examine this.

An exception to the above finding has been the oral text *Robinson Crusoe* used in the November 2006 listening test (see Appendix B, tables 3a and 3b). Thus, although the questionnaire results reveal that 70.8% of the respondents consider the text to be particularly demanding due to lexical difficulty (60.8%) and topic unfamiliarity (59.7%), item analysis indicates that only two items (i.e., item 12 and 20) out of the ten addressing the particular text caused great difficulty to the candidates. The paradox in this finding lies with the fact that a great number of respondents (59.7%) are found to be unfamiliar with the story of Robinson Crusoe. Indeed, it was really unexpected that neither the younger nor the older candidates who have sat for the exam have heard this story before.

A similar conclusion can be drawn as regards the respondents' perceived lack of background knowledge in terms of the oral text *Aesop's fable* (see Appendix B, table 4b). Provided that this text originates from the Greek culture, it seems awkward that 54% of the respondents claim to have little background knowledge about the text. However, this finding is shown to be consistent with item difficulty (see Appendix B, table 4a). Based on my examination of the salient characteristics of the particular text, the use of low frequency vocabulary (*greed, envy, vices etc*.) not only confirms the high percentages of lexical difficulty (58.3%) as it is evident in Table 4b (see Appendix B) but also seems to partly account for the consistency with item difficulty.

Comparing the results from tables 4a and 4b (see Appendix B), another case of inconsistency between item difficulty and text difficulty is noted. While the text USA's *Multilingualism* used in activity 2 is considered difficult by a great number of respondents, item analysis indicates the exact opposite, namely, that candidates did not face any particular difficulty in responding correctly. Here, the respondents' perceptions of text difficulty are mainly attributed to speaker's accent (66.3%), lexical difficulty (54.8%) and lack of interest in the topic (60.5%).

In terms of text interest, the finding can be justified by the specialized topic of the text (i.e., USA's multilingualism), which seems to address the interests, knowledge and experiences of a specific group of candidates (i.e., older, educated candidates). In addition, the specialized topic can partly explain why the respondents have rated the text so high in terms of lexical difficulty (54.8%). However, the inconsistency of these findings with the item analysis data from Table 4a (see Appendix B) implies that the candidates have probably found the test items associated with this text less difficult.

Regarding speaker's accent, the inconsistency with item difficulty can generally be explained by the high percentages of difficulty evident in Table 4b (see Appendix B) in terms of the other texts as well (*Brief oral messages* – 60.6%, *USA's multilingualism* – 66.3% and *Aesop's fable* – 70.8%). It can therefore be assumed that the candidates tend to face difficulties in understanding oral speech probably because they are not as familiar with authentic spoken language as it was expected. This can be explained by taking into consideration the fact that L2 learners are rarely exposed to authentic listening situations in the learning classroom whereas the KPG listening test is designed mainly on the basis of authentic or semi-authentic material.

On the other hand, the low number of difficult test items in Table 4a (see Appendix B) suggests that candidates are not as influenced as they believe by their inadequacy to understand everything they are listening to. This finding supports Rost's (1990) view that second language learners cannot keep up with the language when it is spoken in normal speed and they feel that if they had more time to think about what they are hearing, they would have much less trouble understanding.

Another interesting finding that deserves our attention concerns the role of anxiety in listening comprehension performance. As results from Tables 2b and 3b (see Appendix B) show, the respondents' levels of anxiety are raised according to the difficulty level of the oral texts. Most importantly, consistency is observed between this finding and the results derived from item analysis (see Appendix B, tables 2a, 2b and 3a, 3b) (the only exception being the text *Robinson Crusoe*). This can lead to the assumption that anxiety is likely to influence candidates' performance. Though feelings of anxiety are more or less expected in testing situations, the correlation of candidates' levels of anxiety with text difficulty and item difficulty needs further exploration.

4 Discussion

The main finding of this study is that there is a correlation between item difficulty and candidates' perceptions of text difficulty. In particular, candidates' rating of text difficulty in terms of vocabulary use, background knowledge and topic interest has demonstrated the same hierarchical order of difficulty with the test items displaying unacceptable values of difficulty. Such results seem to suggest that candidates' responses to the test items are influenced by their perceptions of text-related difficulties. This conclusion has important implications for the development of listening comprehension tests as it provides useful insights into the factors underlying candidate performance. Determining what it is that

causes difficulty to candidates in responding, will prove useful for item writers to design valid and reliable listening comprehension tests.

Important conclusions are also drawn from the examination of the cases of data inconsistency revealed in the study. The first finding concerns the paradox that high percentages of respondents are shown to claim lack of background knowledge with texts they are expected to be familiar with (the cases of *Robinson Crusoe* and *Aesop's fable* presented in the previous section). This discrepancy seems to provide support to Vandergrift's (2006) conclusion that L2 learners are either unable to transfer inferencing ability from the L1 drawing on nonlinguistic knowledge resources (e.g., world knowledge) or unaware that they are actually doing it. In the *Robinson Crusoe* text, the fact that the high rates of lexical difficulty are found to be inconsistent with item difficulty may imply that the candidates actually made use of their background knowledge to compensate for their lack of linguistic resources (i.e., vocabulary knowledge) without being aware of it.

Obviously, this finding has clear implications for L2 pedagogy. It suggests that learners may benefit from strategy instruction when responding to listening activities. Relevant research on strategy use and language test performance has shown that L2 listeners need to learn to become more reliant on guessing from contextual or prior knowledge in order to compensate for difficulties with processing oral input (Olsen & Huckin, 1990; O'Malley, Chamot & Küpper, 1989; Tsui & Fullilove, 1998). Similar studies of strategy use by L2 learners have indicated that high proficient learners are more successful listeners than low proficient learners because they tend to make connections between what they listen to and what they already know. In contrast, the less proficient learners consistently rely on words, spelling and pronunciation (Bacon, 1992a, 1992b; Murphy, 1985) (cited in Seo, 2005, pp. 64-65). It is, therefore, implied that language teachers will be able to enhance their learners' L2 listening skills if they encourage them to activate strategies related to the use of their prior knowledge (i.e., inferencing, elaborating, etc.).

Inconsistency is also found in candidates' perceptions of speaker's accent as a factor of difficulty in the oral texts, while item analysis indicates the exact opposite, namely, that they performed successfully. Provided that the phonological features of the English language are remarkably different from the Greek language, this inconsistency is to some extent justified. However, as the data reveal, test performance is not influenced to the extent candidates believe. This finding seems to support several views found in the literature that listening comprehension is very difficult for L2 learners due to distinctive features (i.e., speech rate, accent, elision, the placement of stress and intonation, redundancy and hesitation) that are not found in any of the other language skills (Anderson & Lynch, 1988; Buck, 1991, 1992,

2001; Flowerdew, 1994; Lund, 1991; Rost, 1990, 2002; Ur, 1984). Again important implications are drawn for foreign language teaching and the way listening is taught in the EFL classroom. Clearly, L2 learners should be given more opportunities for exposure to authentic listening situations where the foreign language is spoken naturally.

Finally, the study provides interesting findings regarding the role of anxiety in test performance. Namely, it demonstrates that candidates' levels of anxiety correlate both with text difficulty and item difficulty. Thus, it can be assumed that anxiety may have had an effect on their performance. This conclusion certainly deserves our attention: while it seems natural for candidates to feel anxious in a testing situation, this anxiety must not in any way impede them from responding correctly. This would constitute a threat towards the validity and reliability of the test. There are also implications for pedagogy: the fact that candidates get so anxious during the listening test procedure suggests that they probably feel insecure about their listening abilities. Moreover, it seems necessary that they are provided with more practice in employing test-taking strategies for overcoming any test difficulties.

5 Conclusion

Data were derived from two different sources, namely item analysis and questionnaire analysis. What I actually discovered is that text difficulty in terms of vocabulary use, topic familiarity and topic interest can have an effect on candidate performance. However, background knowledge as a variable of difficulty behaved in a rather unpredictable manner. Thus, a great number of respondents claimed lack of background knowledge with texts whose topics were expected to be generally known to Greek candidates. Obviously, further research is necessary to examine this inconsistency. Moreover, the respondents' perceptions of speaker's accent as a factor of text difficulty were not confirmed by the item analysis data. The authentic or semi-authentic texts used in the KPG listening exam seems to partly account for this inconsistency, given the fact that Greek learners have limited opportunities for authentic listening practice in the learning classroom. Finally, text anxiety was found to influence text difficulty and item difficulty, a finding that certainly deserves further attention.

Unlike previous research focusing on listening comprehension difficulty by analysing either test items or candidate questionnaires, the originality of the present study lies in the fact that it combines the two research methods to achieve its aims. However, lack of evidence in terms of the effect of specific task difficulty variables on candidate performance should be regarded as a limitation of the present study. Further research could build on the current study and look at the influence of those factors on listening comprehension performance.

6 References

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

Anderson, A. & Lynch, T. (1988). *Listening*. Oxford: Oxford University Press.

Arnold, J. (2000). Seeing through listening comprehension exam anxiety. *TESOL Quarterly, 34*(4), 777-786.

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

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

Bacon, S. M. (1992a). Authentic listening in Spanish: how learners adjust their strategies to the difficulty of the input. *Hispania*, *75*, 398-342.

Bacon, S. M. (1992b). Phases of listening to authentic input in Spanish: a descriptive study. *Foreign Language Annals*, *25*, 317-333.

Brindley, G. & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*(4), 369-394.

Buck, G. (1991). The testing of second language listening comprehension: an introspective study. *Language Testing*, *8*(1), 67-91.

Buck, G. (1992). Listening comprehension: construct validity and trait characteristics. *Language Learning*, *42*(3), 313-357.

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

Buck, G., Tatsuoka, K., Kostin, I. & Phelps, M. (1997). The sub-skills of listening: rule-space analysis of a multiple-choice test of second language listening comprehension. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (Eds.), *Current developments and alternatives in language assessment*. Jyväskylä, Tampere: University of Jyväskylä.

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

Common European framework of reference for languages: teaching, learning and assessment (2001). Modern Languages division of the Council of Europe and Cambridge University.

Flowerdew, J. (Ed.). (1994). *Academic listening: research perspectives*. Cambridge: Cambridge University Press.

Freedle, R. & Fellbaum, C. (1987). An exploratory study of the relative difficulty of TOEFL's listening comprehension test items. In R. Freedle and R. Duran (Eds.), *Cognitive and*

linguistic analyses of test performance (pp. 162-192). Norwood, New Jersey: Ablex Publishing Corporation.

Freedle, R., & Kostin, I. (1999). Does text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, 16(1), 2-32.

Graham, S. (2004). Giving up on modern foreign languages? Students' perceptions of learning French. *Modern Language Journal*, *88* (2), 171-191.

Graham, S. (2006). Listening comprehension: the learners' perspective. *System, 34*, 165-182.

Jensen, C., Hansen, C., Green, S. & Akey, T. (1997). An investigation of item difficulty incorporating the structure of listening tests: a hierarchical linear modeling analysis. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 151-64). Jyväskylä, Tampere: University of Jyväskylä.

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

Murphy, J. M. (1985). An investigation into the listening strategies of ESL college students. Doctoral Dissertation, Teachers College, Colombia University, Colombia.

Nissan, S., DeVincenzi, F. & Tang, K. L. (1996). An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension. Research report 51. Princeton, New Jersey: Educational Testing service.

Olsen, L. A. & Hucklin, T. N. (1990). Point-driven understanding in engineering lecture comprehension. *English for Specific Purposes*, *9*, 33-47.

O'Malley, J. M., Chamot, A. U. & Küpper, L. (1989). Listening comprehension strategies in second language acquisition. *Applied Linguistics*, *10*, 418-437.

Purpura, J. E. (1991). *Learner strategy use and performance on language tests: a structural equation modeling approach*. Cambridge: Cambridge University Press.

Rost, M. (1990). *Listening in language learning*. London: Longman.

Rost, M. (2002). *Teaching and researching listening*. Harlow, England: Longman.

Seo, K. (2005). Development of a listening strategy intervention program for adult learners of Japanese. *International Journal of Listening*. Retrieved from http://www.listen.org

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: the effect of text and question types. *Language Testing*, *8*(1), 23-40.

Spelberg, H. C., de Boer, P. & van den Bos, K. P. (2000). Item type comparisons of language comprehension tests. *Language Testing*, *17*(3), 311-322.

Tsui, A. and Fullilove, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics, 19* (4), 432-451.

Ur, P. (1984). *Teaching listening comprehension*. Cambridge: Cambridge University Press.

Vandergrift, L. (2006). Second language listening: listening ability or language proficiency? *The Modern Language Journal*, *90*, 6-18.

Yanagawa, K. & Green, A. (2008). To show or not to show: the effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *Science*, *36*, 107-122.

APPENDIX A

The set of questions about listening comprehension as extracted from the questionnaires[5]

Please answer the following questions by putting a tick ✓

1. I found the oral texts:	**VERY DIFFICULT**	**DIFFICULT**	**EASY**	**VERY EASY**
a. Text A *(title of the oral text)*				
b. Text B				
c. Text C				
d. Text D				

2. I found the vocabulary in the oral texts:	**VERY DIFFICULT**	**DIFFICULT**	**EASY**	**VERY EASY**
a. Text A *(title of the oral text)*				
b. Text B				
c. Text C				
d. Text D				

3. I found the speaker's accent in the oral texts:	**VERY DIFFICULT**	**DIFFICULT**	**EASY**	**VERY EASY**
a. Text A *(title of the oral text)*				
b. Text B				
c. Text C				
d. Text D				

4. I was anxious while listening to the oral texts:	**VERY MUCH**	**MUCH**	**LITTLE**	**NOT AT ALL**
a. Text A *(title of the oral text)*				
b. Text B				
c. Text C				
d. Text D				

5. I liked the topic of the oral texts:	**VERY MUCH**	**MUCH**	**LITTLE**	**NOT AT ALL**
a. Text A *(title of the oral text)*				
b. Text B				
c. Text C				
d. Text D				

6. I was familiar with the topic of the oral texts:	**VERY MUCH**	**MUCH**	**LITTLE**	**NOT AT ALL**
a. Text A *(title of the oral text)*				
b. Text B				
c. Text C				
d. Text D				

[5] The questions were in the form of Likert scales. They were originally provided in the candidates' first language (e.g. Greek) and were translated into English by the researcher for the purposes of the present paper.

7. The quality of sound in the oral texts influenced my understanding:	VERY MUCH	MUCH	LITTLE	NOT AT ALL
a. Text A *(title of the oral text)*				
b. Text B				
c. Text C				
d. Text D				

APPENDIX B

Correlation results[6]

TABLE 2a:May 2006 item analysis data (Total number of test items: 25)

PROBLEMATIC TEST ITEMS							
Activity 1 – 6 M/C test items (*Radio News*)		**Activity 2 – 7 M/C test items** *(Language Museum)*		**Activity 3 – 6 T/F/NS test items** *(South France)*		**Activity 4 – 5 Short answers** *(Movie extract)*	
Item 1 Item 5	0.24 ~ 0.30	Item 10 Item 11 Item 13	0.29 0.27 0.29	Item 16 Item 17 Item 18 Item 19 Item 20	~0.30 <0.40 0.22 0.28 0.29	Item 21 Item 22 Item 23	<0.20 (too difficult) 0.25 <0.20 (too difficult)

TABLE 2b: May 2006 Questionnaire analysis data [frequency]

	Radio News **(Activity 1)**		*Language Museum* **(Activity 2)**		*South France* **(Activity 3)**		*Movie extract* **(Activity 4)**	
	Very Difficult / Very Much	Very Easy/ Little	Very Difficult/ Very Much	Very Easy/ Little	Very Difficult/ Very Much	Very Easy/ Little	Very Difficult / Very Much	Very Easy/ Little
Text Difficulty	**53.2%**	37%	**56%**	34.6%	**60%**	29.8%	**68.4%**	21.4%
Text Lexical Difficulty	43.4%	45.4%	48.2%	42%	**53%**	36.6%	**52.6%**	36.6%
Background Knowledge	41.8%	47.8%	36.6%	**53.2%**	34.2%	**55.2%**	33.2%	**56.4%**
Topic Interest	42.4%	46.8%	45.6%	43%	38%	**51%**	31%	**56.4%**
Text Anxiety	34,9%	44.8%	44.4%	44.8%	**50.8%**	39.2%	**58.4%**	31.2%

[6] The following tables (2a-4b) present the item analysis data and the questionnaire analysis results elicited from the examination periods of May and November 2006 and May 2008.

Overall Task Difficulty	Very Difficult	53.4%	Very Easy	23%

N=500

TABLE 3a: November 2006 item analysis data (Total number of test items: 25)

PROBLEMATIC TEST ITEMS							
Activity 1 – 6 M/C test items *(TV/Radio News)*		**Activity 2 – 4 M/C test items** *(Interview with Grace Kelly)*		**Activity 3 – 4 M/C & 6 T/F/NS test items** *(Robinson Crusoe)*		**Activity 4 – 5 Short answers** *(Notting Hill)*	
Item 1 Item 5 Item 6	0.88 (too easy) 0.33 0.37	Item 7 Item 8 Item 10	0.82 (too easy) 0.85 (too easy) 0.31	Item 12 Item 15 Item 20	0.30 >0.80 (too easy) ~0.30	Item 22 Item 23 Item 24	~0.30 ~0.30 ~0.10 (too difficult)

TABLE 3b: November 2006 Questionnaire analysis data [frequency]

	TV/Radio News **(Activity 1)**		*Interview with Grace Kelly* **(Activity 2)**		*Robinson Crusoe* **(Activity 3)**		*Notting Hill* **(Activity 4)**	
	Very Difficult / Very Much	Very Easy/L ittle	Very Difficult / Very Much	Very Easy/ Little	Very Difficult / Very Much	Very Easy/ Little	Very Difficult/ Very Much	Very Easy/ Little
Text Difficulty	49.2%	45.8%	49.8%	44.6%	**70.8%**	24.7%	**80.3%**	14.8%
Text Lexical Difficulty	24.5%	49.4%	43.1%	**50.6%**	**60.8%**	33.8%	**66.2%**	27.5%
Background Knowledge	39.4%	**55.4%**	37.7%	**56.7%**	34%	**59.7 %**	24.5%	**68.6%**
Topic Interest	**50.6%**	44.6%	**52.4%**	42.1%	46.7%	47.8%	31.6%	**61.7%**
Text Anxiety	44.9%	49.6%	46.7%	47.5%	**56.9%**	37.7%	**68.3%**	26.9%
Overall Task Difficulty	Very Difficult		64.4%		Very Easy		25.7%	

N=494

TABLE 4a: May 2008 item analysis data (Number of test items: 20)

PROBLEMATIC TEST ITEMS		
Activity 1 – 6 M/C test	**Activity 2 – 9 M/C test**	**Activity 3 – 5 Short answers**

items *(Brief oral messages)*		**items** *(USA's multilingualism)*		*(Aesop's fable)*	
Item 1 Item 2	0.39 0.27 (too difficult)	Item 8 Item 15	~0.85 (too easy) 0.26 (too difficult)	Item 16 Item 17 Item 18	<0.20 (too difficult) <0.30 <0.40

TABLE 4b: May 2008 Questionnaire analysis data [frequency]

	Brief oral messages **(Activity 1)**		*USA's multilingualism* **(Activity 2)**		*Aesop's fable* **Activity 3**	
	Very Difficult/ Very Much	Very Easy/ Little	Very Difficult/ Very Much	Very Easy/ Little	Very Difficult/ Very Much	Very Easy/ Little
Speaker's accent	**60.6%**	37.1%	**66.3%**	31.2%	**70.8%**	26.4%
Text Lexical Difficulty	**53.3%**	44.1%	**54.8%**	42.9%	**58.3%**	39.3%
Background Knowledge	47.7%	**49.9%**	45.3%	**52.4%**	43.7%	**54%**
Topic Interest	44.2%	**52.8%**	36.5%	**60.5%**	37.8%	**59.2%**
Sound Quality	**58.4%**	39.5%	**58.8%**	38.7%	**61.2%**	36.4%
Overall Test Difficulty	Very Difficult	74.5%		Very Easy	23.5%	

N= 750