Investigating test taker processes on two ESP listening tasks

Final report

Kellie Frost

Language Testing Research Centre

University of Melbourne

December 2013

Contents

EXECUTIVE SUMMARY

This report describes a qualitative research study designed to investigate the appropriateness of the construct underlying the OET listening test. The purpose of the OET listening test is to provide test users with a valid and reliable assessment of candidates' listening ability in a health-related context. The test consists of two parts: (i) Part A: a note-taking task which requires test takers to keep detailed case notes when listening to a recorded consultation between a patient and a health professional, and (ii) Part B: a recorded mini lecture on a health-related topic which test takers are required to listen to and complete a range of question types as they listen, including multiple-choice questions, short answer questions, gap fill exercises and sentence completions. The tasks of the listening test are designed to reflect the sorts of listening demands health professionals are likely to encounter in the workplace. The two parts of the test, A and B, are intended to draw on different listening skills, thereby capturing a more comprehensive and authentic listening construct.

In the current study, a verbal reporting methodology was used to investigate if the listening processes and behaviours that test-takers engage in on the test resemble those that the test tasks are designed to elicit, and to verify that the two parts of the test are tapping into different aspects of listening skills, as expected. In other words, the aim was to evaluate if the hypothesised construct, articulated in the test specifications, is in fact being operationalized by the test tasks encountered by candidates. Qualitative research of this nature, which can shed light on test taker knowledge, processes and strategies, is required to supplement traditional score-based statistical analyses in order to support claims of validity. In terms of building an overall validity argument for the OET, the study provides construct-related evidence in support of the explanation inference, one of the six principal inferences underlying the interpretation of test scores in an argument-based approach to test validation (Chapelle, Enright and Jamieson, 2010; Kane, 1992; Kane, Crooks & Cohen, 1999; Xi, 2008). The explanation inference relies on the assumption that the listening processes, skills and strategies elicited by the tasks are consistent with theoretical expectations.

This project had the further practical aim of refining where necessary the design of the listening tasks to ensure their appropriateness for providing evidence in support of conclusions about candidates' listening ability and language readiness for work in health-related contexts.

The research questions addressed in the current study were as follows:

1. What listening skills and strategies do the tasks and items on the OET listening sub-test measure?

2. Are there differences between the types of skills and strategies engaged by test takers in response to Part A of the test compared to Part B?

The report begins with a review of the literature concerning listening comprehension and assessment. An overview of current theoretical models of listening comprehension is first presented, with specific attention paid to descriptions of cognitive processes, listening sub-skills and listening strategies, before turning to a review of literature related specifically to the assessment of listening comprehension. The characteristics of test takers who volunteered for the study are then detailed. Following this, the methodology is outlined, which involved 30 test takers completing shortened versions of both parts of the listening test, and providing verbal reports at the end of each question of each part of the test. Specifically, they were asked to recount their thought processes as they completed each question of Part A and Part B, and to comment on any areas of difficulty.

Results suggest that there is strong evidence to support the taxonomy of abilities in the existing test specifications, and provide evidence that many of these abilities are important for distinguishing between test takers with different levels of listening proficiency. Clear qualitative differences were observed between the listening abilities and range of strategy use of weaker versus stronger participants on both parts of the test. Further, the data provide strong evidence that the two parts of the OET listening test made different and appropriate demands on test takers. Evidence also indicates that the different task types in Part B create a broader range of difficulty on the test, distinguishing more finely between weaker participants. This study thereby provides valuable evidence in support of the overall validation argument for the OET.

Some potential issues and areas for improvement were identified, which led to the following recommendations:

- Clarification of the specifications in relation to formatting guidelines
- Clarification of instructions to test takers where problems have been identified
- Development of more specific instructions for item writers in terms of achieving a broad and measured spread of item difficulty, through insights obtained into the relationship between text features, item-type and difficulty
- Further investigation of potential sources of construct-irrelevant variance, such as the impact of written summary skills, reading ability and spelling knowledge

INTRODUCTION

The Occupational English Test is a specific purpose test designed to evaluate the English-language competence of qualified medical and health professionals who wish to practise in an English-language context. It seeks to ensure that candidates are prepared, in language terms, for work in their profession. The purpose of the OET listening sub-test is to provide test users with a valid and reliable assessment of candidates' listening ability in a health-related context. The listening sub-test consists of two parts: (i) Part A: a note-taking task which requires test takers to keep detailed case notes when listening to a recorded consultation between a patient and a health professional, and (ii) Part B: test takers are required to listen to a recorded mini lecture on a health-related topic and complete a range of question types, including multiple-choice questions, short answer questions, and summary completions.

The tasks of the listening sub-test are designed to reflect the sorts of listening demands health professionals are likely to encounter in the workplace. The two tasks are intended to draw on different listening skills, thereby capturing a more comprehensive and authentic listening construct. The test specifications, detailed below, indicate abilities that test takers are hypothesized to draw on when completing each of these two task types.

Broadly speaking, Part A of the OET listening subtest assesses a candidate's ability to "follow the facts" during a consultation between a health professional and a patient as evidenced through a note-taking task (McNamara, 1990: 204). Following the facts involves understanding content, extracting relevant information, and taking note of relevant details while listening to a consultation in real time.

A number of specific listening abilities are hypothesized to be tested through the note-taking task. According to the current test specifications, these include (drawing on Richards, 1983):

- The ability to discriminate between the distinctive sounds of the target language
- The ability to recognize the functions of stress and intonation to signal the information structure of utterances
- The ability to identify words in stressed and unstressed situations
- The ability to recognize reduced forms of words
- The ability to distinguish word boundaries
- The ability to understand vocabulary (general, technical and colloquial)

- The ability to detect key words and phrases
- The ability to guess the meaning of novel lexical items from the contexts in which they occur
- The ability to recognize syntactic patterns and devices
- The ability to recognize cohesive devices in spoken discourse
- The ability to recognize elliptical forms of grammatical units and sentences
- The ability to detect meanings expressed in differing grammatical forms/sentence types (i.e., that a particular meaning may be expressed in different ways)
- The ability to recognize the communicative functions of utterances, according to the context of the consultation, the participants and their goals
- The ability to infer links and connections between events
- The ability to deduce causes and effects from explicitly described events
- The ability to recognize coherence in discourse, and to detect such relations as main idea, supporting idea, given information, new information, generalization, exemplification
- The ability to process speech at different rates
- The ability to process speech containing pauses, errors and corrections

Part B of the OET listening subtest assesses candidates' ability to understand a short talk on a health-related topic that might realistically occur in a health-related work context. According to the test specifications, listening comprehension on this task will involve:

- Direct meaning comprehension
    - Understanding an overarching argument or point
    - Understanding main ideas and important information
    - Listening for specific details
    - Understanding health-specific vocabulary

- Inferred meaning comprehension
    - Making inferences from information available in the text
    - Inferring meaning of unfamiliar lexical items from context
    - Recognising the communicative functions of utterances, according to the context of talk, the speaker and his/her goals

The listening sub-skills that this task aims to engage include (drawing on Richards, 1983):

- The ability to process speech at different rates, from different age groups, and with different accents
- The ability to process speech containing pauses, errors and corrections
- The ability to recognize the functions of stress and intonation to signal the information structure of utterances
- The ability to understand vocabulary (general and colloquial)
- The ability to recognize cohesive devices in spoken discourse
- The ability to detect meanings expressed in differing grammatical forms/sentence types (i.e., that a particular meaning may be expressed in different ways)
- The ability to infer links and connections between events
- The ability to deduce causes and effects from explicitly described events

As can be seen, the abilities listed for Part B in the test specifications overlap with those listed for Part A. The different question types of Part B are intended to achieve a more nuanced spread of difficulty across the test by allowing specific abilities to be more directly targeted. The aim of the current project was to use verbal report methods to explore the construct validity of the two listening tasks by investigating if the processes and sub-skills test-takers engage in resemble those which the tasks are designed to elicit, as defined in the test specifications, cited above. The project provides insights into the listening processes, skills and strategies engaged by each of the two tasks. Findings provide evidence to support the claim that the two tasks are measuring different aspects of the listening construct, and also that the tasks are effectively measuring the overall intended construct.

Our project had the practical aim of refining where necessary the design of the listening tasks to ensure their appropriateness for providing evidence in support of conclusions about candidates' listening ability and language readiness for work in health-related contexts. While evidence from the project support current task design, this report concludes with a recommendation to update the test specifications in order to include more detailed descriptions of task-specific listening demands.

LITERATURE REVIEW

Introduction

Qualitative research which can shed light on test taker knowledge, processes and strategies is required to supplement traditional score-based statistical analyses in order to support claims of construct validity. Verbal reporting is an established methodology that serves this purpose. Gass and Mackey (2000) define verbal reporting 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). Although verbal reports have been used within second language learning research more widely to investigate other language skills, a growing body of research has utilized the method in studies related to second-language listening comprehension. Some of the features of listening which have been investigated include: strategy use (e.g., Goh, 1998; O'Malley, Chamot & Kupper, 1989; Vandergrift, 2003), the nature of listening difficulty (e.g. Goh, 2000), and the utilisation of visual information in video texts (e.g., Gruba, 1999; Ockey, 2007; Wagner, 2008). The method has been applied in other testing-related research (see Lumley & Brown, 2005), and Green (1998) suggests that verbal reports may be used for a range of validation purposes. With respect to listening assessment, two prominent studies by Buck (1991) and Wu (1998) have drawn on verbal report methods to investigate the processes of test-takers on listening assessment tasks. For existing testing programs, verbal reports can be used to investigate whether the hypothesised construct, articulated in the test specifications, is in fact being operationalized by the test tasks which candidates encounter.

In the current project, the aim was to use verbal reports to gather construct-related evidence in order to evaluate the explanation inference, one of the six principal inferences underlying the interpretation of test scores in an argument-based approach to test validation (Chapelle, Enright and Jamieson, 2010; Kane, 1992; Kane, Crooks & Cohen, 1999; Xi, 2008). The explanation inference relies on the assumption that the listening processes, skills and strategies elicited by the tasks are consistent with theoretical expectations, as articulated in the test specifications.

In light of this aim, a review of the literature concerning listening comprehension and assessment is given below. The review begins with an overview of current theoretical models of listening comprehension, with specific attention paid to descriptions of cognitive processes, listening sub-skills and listening strategies, before turning to a review of literature related specifically to the assessment of listening comprehension.

Models of listening comprehension

Broadly speaking, the cognitive processes, skills and strategies involved in second language listening are thought to be similar to those involved in first language listening (Buck, 2001; Flowerdew & Miller, 2005) and as a consequence, first language comprehension models have provided the basis for theoretical conceptualisations of second language listening comprehension.

One influential model which provides an account of the way in which auditory messages are attended to and processed in first language contexts is the Human Information-Processing System model (Bourne, Dominowski & Loftus, 1979). The model involves three types of memory stores. The first is the sensory memory store in which auditory messages from the environment are detected and held intact for no more than one second. Then, depending on the nature of the message (relevance, importance, etc), it is either transferred to the short-term memory or discarded. In short-term memory the message is held for no more than 15 seconds and is subject to conscious processing in which information is classified as new or old. Old information is checked against information held in long-term memory stores, and listeners attempt to make sense of new information by matching it against existing knowledge held in long-term memory. This then allows the new information to be stored in long term memory where it can be categorised, assessed and fully interpreted (Flowerdew & Miller, 2005).

In terms of second language processing, Buck (2001) refers to a model proposed by Nagle and Sanders (1986) in which three types of memory are similarly distinguished. The first is echoic memory, which is much the same as sensory memory, above; the second is working memory, consistent with short-term memory, above, except that here the distinction between controlled and automatic processing is highlighted as particularly relevant in second language listening. As Buck explains, auditory input is "processed in working memory by an executive processor, by means of either controlled processes or automatic processes, or any degree of combination between the two, and the result is then passed to long term memory" (2001: 27). The third type is long term memory, as also specified in the Human Information-Processing System model.

Buck (2001) argues that while the model proposed by Nagle and Sanders (1986) is useful for understanding second language listening comprehension, it fails to explain how text meaning is built up in memory. He draws on van Dijk and Kintsch's (1983) model of comprehension to explain how listeners develop a mental model of text meaning. According to Van Dijk and Kintsch (1983), comprehension involves both the construction of a "textbase" or semantic representation of the text, whereby information from each proposition is combined and integrated with the meaning elements from previous propositions in a continuing and iterative process; and the construction of a "situation model", in which interpretation involves the combination of background (world and contextual) knowledge with text meaning. As Buck summarises, "the listener is creating and updating a mental model while listening, and at any point during the listening process that mental model provides the context for the interpretation of the next part of the text" (2001: 28).

Buck (2001) notes, however, that while theoretical models provide useful, simple metaphors for conceptualising listening, listening comprehension is a complex and multifaceted phenomenon comprised of various interacting cognitive processes, sub-skills and strategies. The ways in which the cognitive processes, sub-skills and strategies involved in listening comprehension have been represented in the literature, particularly in relation to second language listening comprehension, are reviewed below.

Cognitive processes

In terms of models of cognitive processes, Flowerdew and Miller (2005) identify three of the most well-known: (i) The bottom-up model; (ii) the top-down model; and (iii) the interactive model. The bottom-up model describes listening comprehension as a process which begins with individual sounds or phonemes, which are combined into words and eventually sentences. Sentences are then combined "to create ideas and concepts and relationships between them" (Flowerdew & Miller, 2005: 24). By contrast, the top-down model privileges the role of prior background and contextual knowledge which is applied in order to predict meaning and thereby interpret and comprehend utterances. The interactive model, as the name suggests, describes listening comprehension in terms of an interaction between bottom-up and top-down processes, which are both drawn upon in parallel and in different ways by different listeners.

According to Flowerdew and Miller, a model of listening processes must account for four main types of knowledge that may be drawn upon in order to achieve comprehension: "*phonological* - the sound system; *syntactic* - how words are put together; *semantic* - word and propositional knowledge; and *pragmatic* - the meaning of utterances in particular situations" (2005: 30). Buck

(2001) makes a broader distinction between linguistic and non-linguistic knowledge. Linguistic knowledge involves (but is not limited to) knowledge of “phonology, lexis, syntax, semantics and discourse structure”, while non-linguistic knowledge includes “knowledge about the topic, about the context, and general knowledge about the world and how it works” (Buck, 2001: 2). Both types of knowledge are used in comprehension, as noted above, but the latter is privileged in the top-down model of comprehension, while the former is drawn on first according to the bottom-up view.

While acknowledging that there are likely to be similarities in listening comprehension processes for first and second language listeners, Flowerdew and Miller (2005) highlight key differences which may impact the way the comprehension processes outlined in the models above will be engaged in second language contexts. First of all, second language listeners face additional barriers to comprehension, such as limited linguistic, cultural and/or background knowledge which may reduce their ability to compensate for interference, such as that caused by background noise, for example. Furthermore, once a message is in short-term memory first language listeners are able to access automatic processing devices, which allow for fast and efficient decision making about whether the information needs to be stored and attended to further in long term memory. Second language listeners, by contrast, may have to rely more heavily on “controlled processing, which requires more attention before any decision on the message can be made” (Flowerdew & Miller, 2005: 28). This is similar to a point raised by Buck (2001) concerning the speed and real-time nature of spoken language. He suggests that the speed and complexity of normal speech means that automatic processing is needed for full comprehension, and given that second language learners may lack the knowledge and experience needed for automatic processing, they may have insufficient time to process the entire message. As a consequence they may attend more to linguistic features and less to overall interpretation, or they may miss parts of the text. At some speeds, he points out, “their processing will tend to break down completely, and they will fail to understand much at all” (Buck, 2001: 7). Finally, according to Flowerdew and Miller (2005), once a message has been stored in long term memory, second language listeners may possess limited schemata compared to first language listeners, which might lead to poor categorisation of information (the message may be filed in the wrong place), thereby impeding their ability to retrieve information from long term memory, and to effectively match new information with existing information.

Listening sub-skills

Buck (2001) reviews several different approaches to describing the sub-skills that underlie the processes detailed above, including: (i) *The two stage view* in which listening is described in terms of two consecutive stages, the first being extraction of linguistic information which is followed by a stage involving applying the extracted information to a communicative context; (ii*) A cognitive skills approach* whereby listening comprehension ability is viewed in terms of "a series of increasingly complex cognitive skills that can be used to show increasing facility with listening comprehension" (Buck: 2001, 53); and (iii) *Communicative approaches* which extend the two stage view by specifying various skills needed for utilising linguistic information for communicative purposes, including skills needed for direct-meaning, inferred meaning and contributory meaning comprehension, as well as for listening and taking notes. Buck also cites Richards (1983), who proposes a more complete taxonomy of listening sub-skills in order to account for listening comprehension. Richards (1983) argues that different listening purposes require the engagement of different listening sub-skills or micro-skills, and his proposed taxonomy is categorised according to two broad purposes, conversational and academic listening. The detailed taxonomy developed by Richards (1983), reproduced below, served as the basis for deriving the OET listening test construct definition, as described in the specifications noted earlier.

Table 1. Micro-skills: conversational listening (Richards, 1983: 228-229)

1. Ability to retain chunks of language of different lengths for short periods
2. Ability to discriminate among the distinctive sounds of the target language
3. Ability to recognise the stress patterns of words
4. Ability to recognise the rhythmic structure of English
5. Ability to recognise the functions of stress and intonation to signal the information structure of utterances
6. Ability to identify words in stressed and unstressed positions
7. Ability to recognise reduced forms of words
8. Ability to distinguish word boundaries
9. Ability to recognise typical word order patterns in the target language
10. Ability to recognise vocabulary used in core conversational topics
11. Ability to detect key words (i.e., those which identify topics and propositions
12. Ability to guess the meaning of words from the contexts in which they occur
13. Ability to recognise grammatical word classes (parts of speech)
14. Ability to recognise major syntactic patterns and devices
15. Ability to recognise cohesive devices in spoken discourse
16. Ability to recognise elliptical forms of grammatical units and sentences
17. Ability to detect sentence constituents
18. Ability to distinguish between major and minor constituents
19. Ability to detect meanings expressed in differing grammatical forms/sentence types (i.e., that a particular meaning may be expressed in different ways)
20. Ability to recognise the communicative functions of utterances, according to situations, participants, goals
21. Ability to reconstruct or infer situations, goals, participants, procedures
22. Ability to use real world knowledge and experience to work out purposes, goals, settings, procedures
23. Ability to predict outcomes from events described
24. Ability to infer links and connections between events
25. Ability to deduce causes and effects from events
26. Ability to distinguish between literal and implied meanings
27. Ability to identify and reconstruct topics and coherent structure from ongoing discourse involving two or more speakers
28. Ability to recognise markers of coherence in discourse, and to detect such relations as main idea, supporting idea, given information, new information, generalization, exemplification
29. Ability to process speech at different rates
30. Ability to process speech containing pauses, errors, corrections
31. Ability to make use of facial, paralinguistic, and other clues to work out meanings
32. Ability to adjust listening strategies to different kinds of listener purposes or goals
33. Ability to signal comprehension or lack of comprehension, verbally and non-verbally

Table 2. Micro-skills: Academic listening (listening to lectures) (Richards, 1983: 229-230)

1. Ability to identify purpose and score of lecture
2. Ability to identify topic of lecture and follow topic development
3. Ability to define relationships among units within discourse (e.g., major ideas, generalizations, hypotheses, supporting ideas, examples)
4. Ability to identify role of discourse markers in signalling structure of a lecture (e.g., conjunctions, adverbs, gambits, routines)
5. Ability to infer relationships (e.g., cause, effect, conclusion)
6. Ability to recognise key lexical items related to subject/topic
7. Ability to deduce meanings of words from context
8. Ability to recognise markers of cohesion
9. Ability to recognise function of intonation to signal information structure (e.g., pitch, volume, pace, key)
10. Ability to detect attitude of speaker toward subject matter
11. Ability to follow different modes of lecturing: spoken, audio, adio-visual
12. Ability to follow lecture despite differences in accent and speed
13. Familiarity with different styles of lecturing: formal, conversational, read, unplanned
14. Familiarity with different registers: written versus colloquial
15. Ability to recognise irrelevant matter: jokes, digressions, meanderings
16. Ability to recognise functions of non-verbal cures as markers of emphasis and attitude
17. Knowledge of classroom conventions (e.g., turn taking, clarification requests
18. Ability to recognise instructional/learner tasks (e.g., warnings, suggestions, recommendations, advice, instructions)

Listening strategies

The distinction between skills and strategies is not clearly maintained in the literature, and often the two terms are used to refer to the same aspects of listening comprehension processes. Vandergrift (1997), for example, provides a taxonomy of cognitive and metacognitive listening comprehension strategies. Although much less comprehensive, his list of cognitive strategies overlaps with many of the micro-skills listed by Richards (1983), above. For the purposes of this report, listening sub-skills and listening strategies have been categorised separately simply because many of the empirical studies referred to below use the term strategy rather than sub-skill, although it is noted that there is little if any substantive difference between the two terms.

Vandergrift's (1997) distinction between cognitive and metacognitive listening comprehension strategies is drawn from the more general distinction posed by O'Malley and Chamot (1990) between metacognitive and cognitive language learning strategies. According to Vandergrift, cognitive listening comprehension strategies include:

(i) inferencing, or an ability to infer meaning of unfamiliar words from context and to fill in information that is missing. He distinguishes four sub-categories of inferencing: linguistic (use of known words to infer meaning), voice (use of tone or other paralinguistic clues to infer meaning), extralinguistic (use of contextual information outside of the text to infer meaning) and between-part inferencing (use of information from the text beyond individual sentence level to infer meaning);

(ii) elaboration, or the ability to combine background knowledge with text information to fill in information gaps. Here he distinguishes personal, world, academic, questioning and creative elaboration. Personal, world and academic refer to personal knowledge, world knowledge and academic knowledge, respectively. Questioning elaboration refers to an ability to combine questions and world knowledge to come up with logical possibilities, and creative elaboration refers to an ability to create a storyline;

(iii) Imagery, or an ability to use visuals (mental or actual) to represent information;

(iv) summarization, either mental or in written, of the listening information.

Metacognitive listening comprehension strategies include:

(i) planning (advance organization, directed and selective attention, self-management);

(ii) monitoring (comprehension and double-check monitoring);

(iii) evaluation, or checking the outcomes of comprehension for completeness and accuracy;

(iv) problem identification.

In a later article, Vandergrift (2003) compares strategy use by skilled and less skilled second language listeners, finding significant quantitative and qualitative (via think-aloud protocols) differences in the frequency of strategy use by each group, particularly the use of metacognitive strategies, as well as individual differences in the use of some cognitive strategies. On the whole, skilled listeners used metacognitive strategies, primarily comprehension monitoring, more frequently and effectively, and were able to interact more deeply with the text than their less skilled counterparts. Of interest in relation to the current project, he also concluded that strategy use was task dependent.

In a review of the literature on listening comprehension strategies, Berne (2008) draws on several earlier studies into differences in strategy use by more and less proficient listeners and reports conclusions consistent with the findings of Vandergrift (2003). She summarises findings from eight researchers over two decades of studies (1980s and 1990s) to conclude that more proficient listeners use a wider range of strategies more frequently and interactively than less proficient listeners. She also suggests that less proficient listeners engage heavily in bottom up processing, rely on translation and key words as strategies, and are less able to make inferences and verify their assumptions than proficient listeners.

Berne (2008) also highlights differences in the types of cues attended to by first and second language listeners. Drawing on earlier studies by Conrad (1981, 1985) ad Harley (2000), she notes that research has shown that learners of English direct attention to syntactic cues and prosodic cues to interpret unknown words or ambiguous sentences, whereas native speakers are more likely to use semantic rather than syntactic cues (native speakers also attend to prosodic cues). As learners of English acquire more advanced proficiency, however, they begin to make more use of semantic cues.

Berne (2008) further notes that research has suggested that language learners follow similar patterns and sequences of listening strategy use. Citing studies conducted by Martin (1982) and Young (1997), Berne (2008) claims that listening comprehension is an active process in which learners orient themselves to stimulus using contextual or acoustic cues to guess topic, access background and contextual knowledge relevant to the topic as they listen, and during listening actively monitor and evaluate strategy use.

Verbal report studies in listening assessment

As mentioned in the introduction to the literature review section, early investigations by Buck (1991) and Wu (1998) examined the processes of test-takers on listening assessment tasks using verbal report methodologies, and are thereby of particular relevance to the current project. A recent paper by Song (2012) will also be reviewed in this section. Song (2012) investigated the relationship between note taking quality and listening test performance on an English for Academic Purposes test. The study is quantitative but offers useful conclusions about the way in which the note-taking task defines the listening construct, of interest and relevance to the current project.

Buck (1991) used verbal reports to investigate test method effect of short-answer item-types, and to examine if test items can measure higher level cognitive processes, such as inferencing, as well as listeners ability to monitor the appropriateness of interpretation. He also investigated how question preview influenced comprehension and test performance. Buck reported that although short-answer open ended comprehension questions were shown to produce minimal and non significant method effect in previous quantitative analyses (Buck, 1989; 1990), verbal report interviews revealed potential issues with the item-type.

He found, for example, that if the amount of information required in the response is not made clear in the question, then test takers can be unsure of how much information to include. As a consequence, they may write all that they hear and as a result, run out of time and fail to answer subsequent questions. He also highlights problems encountered at the marking stage where decisions have to be made about which responses are acceptable and which are not. Since decisions about the relevance of information can be different between individual test takers as well as assessors, some decisions are likely to be arbitrary. Such arbitrariness impacts the construct definition of a test and therefore test validity. Although a marking guide is carefully developed and consistently applied in the marking of the OET listening sub-test in order to capture a wide range of possible interpretations while ensuring measurement reliability, it remains a potential source of construct irrelevant variance and is thus an ongoing consideration.

As noted in the section above on listening strategies, research indicates that a greater capacity to make inferences and to monitor comprehension distinguishes more proficient from less proficient listeners. In view of the importance of these strategies (or skills), Buck (1991) first examined whether or not test items can measure test takers' ability to make inferences. He found that while some test items had been designed to elicit lower level processing (by asking for information directly from the text) and others were designed to require test takers to make inferences based on information given in the text (higher level processing), "the same item could be testing the ability to understand clearly stated information for one testee and inferencing ability for another" (1991: 76). He also highlights the relationship between inferencing and background knowledge, arguing that differences in cultural assumptions and background knowledge between test takers from different cultural backgrounds "could lead to cases where it is not clear whether the testee had simply not understood and was guessing wildly, or really had understood but had reached different yet perfectly reasonable conclusions" (Buck, 1991: 79).

In relation to comprehension monitoring, Buck claims that despite its importance in listening comprehension, particularly second language listening comprehension "where linguistic knowledge and processing efficiency may be grossly inadequate and the listener is often trying to interpret a text from a partial analysis of the propositional content" (1991: 80), it is unclear if and how this could be tested. Finally, and not surprisingly, he found that question preview (reading the questions before listening) enhanced comprehension, serving to direct decisions about listening purpose and helping test takers predict content.

Wu (1998) used verbal reports to investigate how test takers engaged linguistic and non-linguistic knowledge as they completed a multiple choice listening comprehension task, finding that the type of task defined listening purpose and constrained listening processes. In particular, Wu found that test questions and response options served to activate participants' non-linguistic knowledge, which tended to override information abstracted through linguistic processing if such processing could only be partially accomplished, at times leading to miscomprehension and guessing by less advanced listeners. Accordingly, Wu argues that the multiple choice format favoured more advanced listeners, who were able to achieve full linguistic processing, as they did not need to compensate with the use of background knowledge and beliefs.

Finally, in a recent paper, Song (2012) considers how well note-taking tasks measure listening proficiency compared to open-ended questions. While Song argues that the note taking task is a good indicator of listening proficiency as it allows test takers to demonstrate whatever they understand, on the other hand test takers "might be inclined to take notes of whatever they want to, including details, even if they do not clearly understand the interconnectedness among the ideas" (2012: 83). It is argued that this is a potential issue for note taking tasks in a blank format as opposed to note taking within a supplied outline. The OET note-taking task is somewhere in between the two formats considered by Song, as topic headings are provided to direct test takers towards noting down relevant information. For the purposes of the current project, consideration will be given to whether or not the headings provided achieve this intention, i.e., if test takers report selecting and recording information of relevance, or if they are simply recording all details that they hear without necessarily fully comprehending the text.

To conclude briefly, it is clear that there is much to be gained from the use of a verbal reporting methodology to investigate the construct validity and the appropriateness of the use of the two types of listening tasks and various item-types as measures of OET candidates' listening ability and readiness for work in health related contexts. The literature presented above, as well as the OET test specifications, informed the analysis and interpretation of data in the current project.

METHODOLOGY

Research Questions

In order to provide evidence of the construct validity of the OET listening test and the appropriateness of using two different listening tasks and various item-types, this study drew on verbal report methodology to address the following questions:

1. What listening skills and strategies do the tasks and items on the OET listening sub-test measure?

2. Are there differences between the types of skills and strategies engaged by test takers in response to Part A of the test compared to Part B?

Participants

30 adult participants, for whom English is a second language, were recruited for this study, of which 18 were female and 12 were male. All participants have a health-related professional background. One participant had a background in pharmacy, 10 had a background in nursing, 10 had a background in dentistry, and 6 had a background in medicine. The remaining 3 participants were students pursuing degrees in a health profession. All participants were actual test takers preparing for the test, and had already registered to sit the next administration of the OET, thus ensuring participants are representative of the OET test taker population.

Instruments

The listening sub-test

Participants were asked to complete a previously unseen retired version of the OET listening sub-test, as provided by the OET Centre. The Part A version used was “Chris and the Occupational Therapist”, and the part B version was “Menopause Management”. As detailed in the introduction, the test involves two parts, Part A and Part B. Participants were asked to complete shortened versions of both parts of the test, involving the first five questions (out of ten) for each. The first question of each part of the sub-test is provided as an example, so participants were asked to complete and report on four questions in each part (eight in total). The audio files for both parts of the test had been edited to remove input related to questions six onwards. Test instructions were not altered and participants were expected to follow the same instructions that apply under live test conditions.

Procedures

Verbal report protocols

For each part of the listening sub-test, the audio was paused at the end of each question and an immediate retrospective “think aloud” session was initiated by the interviewer using the following prompt question: *Tell me what you remember thinking as you answered question x.* Optional follow up prompts included: *What do you recall hearing from that section?*; *How did you arrive at these answers for question X?*; *Did you have any difficulty with question X?/ What particular difficulty did you have?*

A small pilot study was conducted to verify that the length of the amended test parts was appropriate and to determine the effectiveness of the verbal report protocol devised for the study.

The audio input related to each question does not exceed 2 ½ minutes, with three out of four questions involving less than 2 minutes of audio input. It was expected that participants would be able to recall their thoughts in sufficient detail with the help of the test paper and their responses as stimuli, but two versions of verbal report procedures were trialled in a pilot study in order to verify expectations. In the first version, the intended study protocol was followed - the audio was paused after each question and a verbal report was elicited. In the second version, an extra pause was inserted in the middle of the audio related to the fourth question in part A and part B of the sub-test, as the audio for these questions exceeded 2 minutes.

The pilot study involved three participants. Two were asked to report after each question of each part of the test (according to the intended protocol), and one was asked to report at two points during question four of each part of the test, after an extra pause in the middle of the audio related to the question and again at the end of the audio. Results showed that all three participants were able to recall their thoughts in sufficient detail after each question of each subtest, regardless of the extra pausing. Further, the time taken for each participant to complete the test and verbal report protocols was within the one hour allocated. As a consequence, it was decided that the length of the amended test and the verbal report protocol were both appropriate for the actual study.

Data analysis

For Parts A and B of the listening test, the number of correct items recorded by each participant was calculated for each of the four questions. The verbal reports from each participant were recorded and transcribed, and then analyzed qualitatively. Specifically, transcripts were segmented and coded thematically by subtest Part (A and B) and question (Q2-Q5). Emerging themes were described and interpreted in light of the task specifications, cited above in the introduction, and the literature reviewed above.

For the purposes of clarity, in the results section below data are organized into groups depending on the number of correct items achieved by participants. For Part A, verbal report data are described under three group headings: (i) Participants with 12-14 correct items (out of 29); (ii) Participants with 15-20 correct items; and (iii) Participants with 22 or more correct items. For Part B, groups were organised as follows: (i) Participants with 11-15 correct items (out of 28); (ii) Participants with 16-23 correct items; and (iii) Participants with 24 or more correct items.

RESULTS

PART A

In this section, a summary of the number of correct items by participants is first provided. This is followed by a description of the verbal report data, which has been organised into three participant groups according to the number of correct items achieved, as outlined above under the heading 'data analysis'. To conclude this section, a summary of the verbal report results for Part A is provided in which findings are linked to the relevant abilities listed in the specifications for Part A of the listening test.

Summary of number of correct responses for Part A

Table 3. Number of correct items for participants 1-15 in Part A

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
Q2	3	5	4	5	5	4	4	3	3	6	3	3	2	5	7
Q3	5	3	3	7	5	5	3	3	6	4	3	5	6	3	4
Q4	6	4	6	7	5	5	4	4	7	8	5	5	4	5	7
Q5	2	3	2	3	2	4	2	2	3	3	3	3	2	2	4
Total*	16	15	15	22	17	19	13	12	19	21	14	16	14	15	22

Table 4. Number of correct items for participants 16-30 in Part A

	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25	P26	P27	P28	P29	P30
Q2	4	2	4	6	2	2	5	5	4	5	5	3	4	4	4
Q3	5	3	6	7	4	6	7	4	5	6	7	6	5	5	5
Q4	8	7	7	9	4	7	9	9	7	9	7	3	7	7	5
Q5	2	2	2	3	3	3	3	4	4	4	4	6	3	3	2
Total*	19	14	19	25	13	18	24	24	20	24	23	18	19	19	16

*Possible marks (Q2-7, Q3-8, Q4-10, Q5-4, Total=29)

As explained earlier, Part A of the listening test is a note-taking task in which test takers are required to list points relevant to the topic heading for each question, provided in the audio text and on the answer paper. Dot points with a blank space in which to write responses are listed below each heading on the answer paper, and the number of dot points corresponds to the number of possible items listed in the marking guide. Points listed by test takers are marked as correct if they correspond with items listed in the marking guide, whereas incorrect or irrelevant information is ignored. Tables 3 and 4, above, show the number of correct items each participant wrote down for each of the four question headings in the shortened version of Part A used in the current study. As can be seen, 6 participants wrote down between 12 and 14 correct items out of a possible 29 (less than half of the possible items), 17 participants listed between 15 and 20 correct items, and 7 participants listed 22 or more correct items. Accordingly, verbal report findings, reported below, are summarised into three broad participant groups: (i) Participants with 12-14 correct items; (ii) Participants with 15-20 correct items; and (iii) Participants with 22 or more correct items.

Verbal Reports

<u>(i) Participants with 12-14 correct items</u>

In general, participants who listed less than half of the possible items reported experiencing unresolvable difficulties due to unfamiliar vocabulary. For example:

P11: *GP give him a x-ray and x-ray show a, a* ***spoil fracture****, um I don't know I don't understand what is spoil fracture but I just write down fracture and spoil* [audio = spiral fracture]

P13: *Still some word I don't know, that's a problem and uh he went to a physio session and uh, involved in a-a* ***scraping exercise*** *something, I'm not sure* [audio = gripping exercise]

When they encountered unknown or unidentifiable words, participants in this group tended to rely on elaborate strategies involving combining phonetic clues, invention and world knowledge to recreate what they might have heard. Often these participants managed to identify only the initial or final sounds of the unknown word, or the final sounds, as illustrated in the examples above, and tended to insert a known lexical item with similar word initial or final sounds that was also semantically possible according to their existing world knowledge, or could be made possible with the invention of some minor details.

P11, for example, was unable to recognise the words 'scuba diving' and inserts 'skyping or skating' instead, which are possible alternatives because they are potentially hobbies and the question heading is 'Chris's exercise, work and hobbies'. Skating seems a more likely hobby (or form of exercise) than skyping, which is probably why the second guess was made. P11 lists an invented detail, 'running', which is not mentioned in the audio, and infers 'play the flue' (the patient, Chris, says that he plays the flute) but clearly has no idea what 'flue' in the current context might mean, except that it is unlikely to be 'flu' the illness, with which he is obviously familiar:

P11: *after he say he do the, skyping or skating oh yeah and do a lot of running in and he mention he play the flue I don't know what he play the flue mean I I pretty sure its not F L U E what's what's the flue… no so I just write write the flue but I pretty sure it's not the flu like vaccination flu like that yeah*

In the example below, P7 is referring to a section of the audio text where the patient, Chris, is explaining to the occupational therapist how his hand injury occurred. The audio text is as follows:

"Um I was drilling a brick wall and the drill gripped and twisted my arm around, with the drill so the drill stayed in place and the drill pushed my arm around"

P7 is unable to identify the word 'drill', despite its repetition, but manages to pick out the word-initial sounds ('dr'). She infers the word 'drawer' in place of 'drill', and invents details in her recount to create semantic coherence: 'open the drawer'. Opening a drawer is something that someone is likely to do with their hand, and so becomes a possible option here where Chris is providing details of how his hand was injured:

P7: Yes ah, um ah I'm thinking uh the man was uh man was told w- he got this in uh hand injury four months ago and uh n- ha- uh got fracture in his uh hand while he was dr- uh dr- uh open the drawer or something it wa- it got twisted then uh sh- uh he said uh he said he felt a injury and he went to doctor

As with P11, P7 also acknowledges that she has missed something and admits a degree of uncertainty concerning her inference. She has accurately identified the word 'twisted' in the audio and appears confused as to how this action ('twisting' or being 'twisted') is related to the object ('drawer'), but is unable to resolve the problem:

P7: I missed the word 'drawer' uh what's uh he was telling the and one incident from where he got injury like he was twisting the drawer I couldn't get uh twisted down drawer I just think I think that the third point I missed some drawer word

Participants in this group were also more likely than others to miss key details and omit essential information from the points they listed. They often managed to pick out and reproduce a key word, but were unable to understand the surrounding text enough to provide any specific details:

P7: *Ah… yes I think uh I missed uh some information he is talking about like tissue uh so I I didn't ah actually get what he was saying about tissue*

P8: *The part that I can't really it's like um… that part that he was saying like scuba diving and playing like I can't really get that one… …I think yes, I did miss, quite a number of things of what he say*

In terms of extracting relevant information, defined in the test specifications as a key aspect of a candidate's ability to "follow the facts" in a consultation between a health professional and a patient, participants in this group typically reported that they tried to write down everything they heard and understood, rather than trying to take notes based on topic heading relevance. For example:

P11: *in that particular time I really don't have the time to think about either relevant or irrelevant it's not … you really just need to write down everything you can hear and you can understand*

This group of participants also reported experiencing difficulties with accent and voice quality, but so did several participants in the better performing groups, as will be shown later:

P8: *The first part I can't really hear what he's talking about, not very clear to me… …I think his pronunciation not very clear to me or maybe because of the accent be- so I I can't really get what he's trying to tell that's the reason, other than that I'm alright*

P17: *I think it's the, pronunciations yeah they have sometimes they swallow the the like uh the last word or something like that yeah I can't really catch*

Finally, participants reported concentration and attention problems, particularly in relation to the requirement to listen and take notes simultaneously, for example:

P13: Yeah because I'm trying fi-fix the, last sentence yeah so I kind of missed uh, yeah the following up

P17: he said uh uh walking, gardening and something I forget, yeah um I found it difficult if if like they say three things… in one sentence, I write down the first one then second one maybe forget and then write the third one so normally I miss one point

(ii) Participants with 15-20 correct items

Similar to participants in the first group, participants with between 15 and 20 correct items often reported experiencing difficulties due to unfamiliar lexical items, for example:

P2: four months ago he had a fracture, fracture of the fourth, actually I don't know the word I know the finger… and ah will he play something, maybe I don't know the word

P10: his hobbies are scoob-diving, I don't know what scoob is and play something which is important for the right hand

Some participants in this group attributed these sorts of difficulties to their unfamiliarity with the terminology and treatments used in other health professions, as in the extract from P3 below:

*P3: I don't exactly ah clarify… what's ah, the exact exercise ripping riping I don't know what's what is it OK why he in which he give him a soft material to do it forty times and I ah I don't ah I don't have any idea about this type of exercise***… …***not medically re- not medical related exercise or the kind of exercise maybe b-because these are not familiar for me.*

In terms of missing key details, whereas participants in the previous group often managed to list known words without providing any specific details, participants in this group, while still encountering several problems and missing key information, were generally able to elaborate further:

P5: *Yep ah… the ah he got the fracture um on his um right hand and um cause he twist, um he twist his arm or something like that I couldn't um, get details…Ah I can hear the words but, like um maybe um I couldn't write because I don't know how to write out the words so and mixing*

P29: *he started to tell about the incident and I suppose he was playing and gripped the wall and twisted arms … something, it is, something he said about pushing arms. Sorry I missed this part… … Mostly related the part I missed … I can't, I couldn't comprehend some words*

In some cases, these participants were able to combine linguistic and professional knowledge to resolve uncertainty and make inferences about what the speaker intended to say:

P6: *for me was a bit unclear so he said that he had a procedure on his knee so it was very, vague information so I I I just hear ah ah he said ar-'arthroscope' or and something related so I associate mm according to my medical knowledge, arthroscope, arthroscopy of the knee that he was trying to say so…*

Whereas participants in the previous group typically reported that they tried to write down everything they heard and understood, rather than trying to take relevant notes, participants in this group reported recognising the need for note-taking and a strategy of focusing on key words or main points, as shown below:

P1: *Yeah um, 'cause it's like take notes, so I don't have enough time to like walking every day or riding bikes or just like um pick up the key words or put, yeah just time saving… …I just picked the, what I think, what I thought ah um key point key words*

P21: *So I actually put …, I wrote … after that, it doesn't matter so if I start from here to here it is like when he was just talking about the medical history I wrote the main points actually, not even the main points the main words.*

Several participants also reported drawing on professional knowledge in order to select information based on perceptions of relevance to the topic heading, for example:

P6: *they're requesting the obvious, ah specific information in regards of his medical history, so that should be things related to his past so I was thinking of his past history so he doesn't have diabetes he doesn't have high blood pressure so all these are associated to the medical history obviously… that's why I started to write it down*

P24: *The questions said Chris's exercise, work and hobbies so I thought I would need to mention walking and gardening and self-employed um, that involves what kinds of things, his hobby, scuba diving I put. Hobbies sometimes have to do with medical conditions.*

While participants in this group generally recognised the need to select and note down relevant information, as noted and exemplified above, many reported confusion and difficulties concerning how much they needed to write, and at times about whether or not particular information should be considered relevant:

P12: *for example he said he's playing flute, it's important because he's uh I mean that he has injury in his right hand so it's important for him because he's playing flute, I don't know if I should write this or not, is it important to write or not*

P18: *for his hobbies he mainly does scuba diving and he plays the flute to help his right hand yeah but at the same time as a note form, I've just written play the flute, um but I haven't written the right hand so I'm not quite sure if, that is irrelevant information or not*

P24: *Mm, the surgery, I wasn't sure whether I needed to write this down three screws during the surgery, I think three screws were used in surgery.*

Participants also report using prediction strategies to direct attention to relevant information in the audio text, and encountering difficulties due to misguided predictions:

P1: *Um, I think like when I was writing I think, thought maybe ah I, maybe like there are more are coming next but actually it's not like yeah so actually I … needed to, like pick more detail from the previous listening not like waiting for the next like you know yeah*

In general, participants in this group found question 4 of Part A difficult, as the audio text is lengthy and dense with details, thus requiring participants to synthesise and summarize in order to produce a brief and accurate account of the relevant information, as exemplified in the following extract:

P6: *he went to the emergency but the the emergency obviously ah everything seems to be alright and they sent back home so he decided to go his general practitioner on the weekend he had an x-ray ah for his bones ah for his bones fracture then ah then he was referred to the orthopaedic surgeon and he ah decided to operate in him and put a plate and three screws and then he had surgery three months ago and that was… basically, so I was trying to synthesise all the information but it was ah, ah ah ah a big chunk of information so… ah so bit complicated to ah write it down*

Part of question 4 involved writing details of how the patient Chris, sustained his hand injury while drilling against a brick wall. The audio text is as follows (also cited above in relation to difficulties due to unfamiliar vocabulary encountered by the previous group):

"Um I was drilling a brick wall and the drill gripped and twisted my arm around, with the drill so the drill stayed in place and the drill pushed my arm around"

This segment of text proved difficult for many participants here. They seemed to find it difficult to follow and conceptualise possibly because the patient seems to have difficulty explaining the event and does not provide a clear and linear account of what happened:

P12: *Mm, I couldn't understand some of this, uh words in the first part… mm about how he, uh twisted his arm… because he was talking with a doctor and he was showing her, and so I didn't see that and I couldn't understand*

P18: *the the fuzzy part that I couldn't recall is how he got it something with a drill and that screwed into his arm but and I couldn't form a proper sentence to fit into this space given so I've just written whatever ah that came to my head…*

P24: *Ah, I didn't get the whole story how he broke his arm. Like doing brick walls, I couldn't get the picture, not exact story um, just it twisted, I really, I just couldn't picture the thing, maybe that's it, yeah*

Similar to participants in the previous group, participants in this group also mentioned experiencing some difficulties with accent and voice quality, for example:

P6: *I couldn't ah, ah, I couldn't exactly hear what he mentioned in the last part so I was very ah struggling a bit ah what he said in the last part I didn't understand him clearly he said mention what your hobbies are and he said it was very whispery so I was trying to see what I could recall but I didn't recall anything*

Participants in this group also reported concentration and attention problems in relation to the requirement to listen and write notes simultaneously:

P2: *I can't write quickly, while I write this one, the next information come up, sometimes I miss information*

P12: *Yeah he said something here… about his physiotherapy, that what he did that I, I was writing so I couldn't concentrate to, uh hear exactly what he uh saying*

P24: *It was too fast, I just couldn't catch up with writing… … I just didn't have much time to write down, he keeps talking and I couldn't write down everything he said.*

Some participants also mentioned that uncertainty about spelling or grammar caused further distraction and increased the difficulty associated with writing and listening simultaneously:

P2: *Yeah like ah, ah well I try to think the, write the.. make sense, like write the whole sentence and ah and I think the spelling, think grammar so yeah I write this one, next they continue speaking so ah yeah most trouble is I can't ah write quickly, yeah*

P14: *Uh, yeah a little bit uh because mm… some word is uh like uh when I spelling, and it's wasting my time to, ah so I missed, the others, following so… that's the, hardest one*

P21: *It was pretty fast and I was just busy writing scuba diving that is I mean I actually wrong spelling for scuba, so I was more concentrating on that and I forgot listening for this*

(iii) Participants with more than 22 correct items

Participants with 22 or more correct items out of a possible 29 reported encountering fewer unresolvable problems than participants in other groups, and were generally better able to compensate for any comprehension gaps.

As with participants in other groups, these participants also encountered unfamiliar lexical items, but were often able to identify and reproduce unknown words, and also to guess at their meaning using contextual cues and existing linguistic and professional knowledge. For example, P4, below, uses professional knowledge to successfully resolve uncertainty concerning the word ‘bolts’:

P4: *Um yeah ah yes I had like, I was not sure if what he said was ‘bolts’ um in hips the three bolts in hip um and I guess this was arthroscope, yeah so just the beginning was bit um not clear but towards the end it was fine*

I: *Yep, um so how did you arrive at your answers?*

P4: *So um I’m in medical profession so I just guessed that could be bolted in the hip and arthroscope so it’s just with knowledge*

As shown above, participants in the previous group typically decided to omit unknown words even if they had managed to identify word boundaries and reproduce the sounds. Participants in the current group were often able to accurately list relevant details even if they were unable to understand the meaning of the phrase that they had heard, for example:

P15: *he mentioned something about soft plastic material and gripping maybe I’m really not sure about that*

P19: *It was not clear to me what he was saying. But what I remember is he told about a couple of physio session and he there is gripping and soft plastic material… …that is what I have written*

In terms of missing key details, while participants in the first group often missed most details, and participants in the second group, while still encountering several problems and missing key information, were generally able to elaborate and provide the main gist, some participants in this group were able to identify specific uncertainties and to articulate the particular detail they had missed, for example:

P15: *about this previous treatment, he had uh and ah talking about the plate and I'm really not sure if it was two screws or three screws*

As with participants in the second group, these participants understood the need to select and take note of relevant information and reported using a strategy of noting key words and main points rather than writing down all of the details they heard, as illustrated by P22's comments, below:

P22: *he said to the emergency department of the hospital I remove all these words no need for them emergency it is only the hospital so I leaved it as it and then uh uh uh not broken, I stuck and this uh he said some words a full sentence but it mean not broken it was not broken, so I have to write in the real life I have to write he was diagnosed as not broken eh uh there is arm uh his arm is not broken but I don't have the time, so I wr-write down only not broken*

P26 reports using key words from the question heading to direct listening attention:

P26: *Um, the initial thought that came to me is okay we know she is asking him about his exercise, work and hobbies. So the first part would be exercise, the second part would be work and the third part would be hobbies. So he answers that he is going, he tries to exercise often and what he does is walking, bike riding and gardening. Um, so I think that is the part that came straight away so um and then she went on to work … ten questions about work …accounting software. Again the question came up do you write self-employed as an accounting software consultant or do we write … but there is two bits, so one is self-employed, the other is what he does. So this is okay.*

Participants in this group also report drawing on professional knowledge in order to select information based on perceptions of relevance to the topic heading, for example:

P19: *If I am a doctor what important things I will check if I will get a case history of that person. So because I should know what should, when he get his surgery or when he broken his arm so that is why I have written. And was the treatment delayed or it was immediate? It was delayed because it was clear that initially ah they told him that it was not broken but then he went to his doctor and he told that it was a spiral fracture. So these are all important things for a doctor. If you are referring a patient to a doctor for example here they are referring to orthopaedic surgeon it is very important that orthopaedic surgeon should know what was the history of the fracture. That is what I thought when I writing this.*

As with participants in the previous group, these participants also reported finding question 4 difficult due to the large amount of information and the lack of clarity in the patient's description of how the injury occurred. These participants, however, were able to note down most of the ten items correctly, and managed to efficiently summarize the cause of the injury. For example:

P25: *He was drilling brick wall and I was not sure here again how to write it down because I understand that he was drilling brick wall and he twisted his arm around … still I was confused how to write it down so I wrote two sentences.*

P26: *Um, for this part of it it was ah, for me to get these answers was slightly different because he didn't answer these questions directly and I am presuming he is talking to her so just pointing what is injured and she prompts him to say, okay so that is a fracture of the fourth metacarpal… …So he was drilling the wall and then the drill got stuck in the wall and because it is still moving so the drill twisted his arm around and… …he was drilling a brick wall then I thought should I write drilling a brick wall or should I write the fact that the drill twisted his arm around so and I decide to write the second part of it, Which is kind of more specific to why the injury happened.*

Similar to the previous groups, these participants also reported experiencing difficulties due to speech rate and the pressure of having to write and listen at the same time, for example:

P23: *I think the, the rate of speaking was a little bit fast so that I couldn't ah write everything and listen in the same time… …I have to write a lot of information and he keep giving all of the information, relevant information following each other without giving time to write.*

They also reported that uncertainty about spelling was a source of difficulty, but were generally able to resolve problems and still follow the audio text, for example:

P22: *I stuck in bicycle because of thought uh can't remember how we write uh bicycle yeah, I forget it then I write it again after I finish… …so I catch the word scuba diving and then eh uh he said uh something about the flute, I think it's an instrument for the the music, yeah but I don't know the spelling of it, I just write it like this ['floot'] I'm sure it is wrong.*

Summary of results for Part A

The main findings for Part A, reported above, can be summarised into three broad categories: (i) findings related to test taker vocabulary and word knowledge; (ii) findings related to identifying and selecting main points and essential information, including selecting information for topic relevance; and (iii) findings related to an ability to follow 'everyday', unscripted speech and conversation. Each category will be discussed briefly below in relation to abilities listed in the existing test specifications.

(i) Findings related to test taker vocabulary and word knowledge

As discussed above, several participants in the first two groups reported experiencing unresolvable difficulties due to unfamiliar words. By contrast, those in the final group were often able to identify and accurately record unfamiliar words, and were also more likely to use contextual clues and existing linguistic knowledge to guess at the likely meaning of novel lexical items. These findings provide evidence that Part A of the test requires test takers to draw upon the following abilities, as listed in the existing test specifications:

- The ability to identify words in stressed and unstressed situations
- The ability to distinguish word boundaries
- The ability to understand vocabulary (general, technical and colloquial)

- The ability to guess the meaning of novel lexical items from the contexts in which they occur

It was also found that participants with the least number of correct items were often unable to distinguish all of the sounds in unknown words, and only picked out the first or last sounds. Better participants were able to record words that were novel because they were able to distinguish all of the sounds. This evidences the following ability in the existing specifications:

- The ability to discriminate between the distinctive sounds of the target language

(ii) Findings related to identifying and selecting main points and essential information, which includes selecting for topic relevance

Data also show that participants in the final group were better able to identify and select relevant information than other participants, and better able to make inferences about which details were essential to include. Participants with the least number of correct items were typically able to identify particular words as relevant, but could not understand enough of the text to sufficiently elaborate. They also tended to write everything they heard rather than filtering for relevance. Participants in the middle group elaborated to some extent, but were unable to capture all of the essential information. This was particularly the case for the more complex segments of text, such as the segment relating to question 4 of Part A, for example. Participants in the third group, by contrast, were able to synthesize lengthy sections of audio and infer important links between events in order to construct a relevant, short summary sentence as an answer. These findings support the inclusion of the following abilities in the existing specifications:

- The ability to detect key words and phrases
- The ability to infer links and connections between events
- The ability to deduce causes and effects from explicitly described events
- The ability to recognize coherence in discourse, and to detect such relations as main idea, supporting idea, given information, new information, generalization, exemplification

As noted in the results, above, participants in all three groups reported concentration and attention problems, particularly in relation to the requirement to listen and take notes simultaneously, as well as experiencing difficulties due to speech speed, accent, and voice quality. The latter findings highlight the importance of the following abilities, listed in the current specifications:

- The ability to process speech at different rates

- The ability to process speech containing pauses, errors and corrections

The difficulty of listening and writing simultaneously reported by many participants is perhaps worthy of further consideration. Better participants were able to note down most of the items correctly, despite reporting difficulties, possibly suggesting greater automaticity, typical of higher proficiency second language users and therefore a legitimate aspect of the construct. It is possible, however, that they simply possessed better note-taking skills, or better written summarization skills, than other participants. While Part A is a note-taking task, note-taking and written summarization skills are not typically considered as aspects of listening ability, and so are not included in the specifications. It would be interesting to further investigate if these sorts of skills impact performance on the listening test in any significant way.

Finally, many participants across all groups noted spelling or grammatical difficulties as a source of distraction, which they reported sometimes led to missing subsequent information. Spelling and grammatical correctness of written responses are not intended to be part of the test construct and are not referred to in the specifications, so it would also be of interest to investigate further the impact of such 'distractions' on test performance. For the most part, however, the findings for Part A support the taxonomy of abilities in the existing test specifications, and provide evidence that many of these abilities are important for distinguishing between test takers with different levels of listening proficiency.

PART B

As with the results for Part A, in this section a summary of the number of correct items by participants is first provided. A description of the verbal report data follows, which has been organised into three participant groups according to the number of correct items achieved. To conclude this section, a summary of the verbal report results for Part B is provided in which findings are linked to the relevant abilities listed in the specifications for Part B of the listening test.

Summary of number of correct responses for Part B

Table 5. Number of correct items for participants 1-15 in Part B

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
Q2	5	3	4	5	2	3	2	1	2	4	2	5	5	5	4
Q3	6	3	5	6	5	6	4	3	5	5	6	4	4	3	6
Q4	7	3	4	6	4	3	2	2	3	4	4	6	4	4	4
Q5	8	6	9	7	3	7	4	5	5	8	4	8	6	3	8
Total*	26	15	22	24	14	19	12	11	15	21	16	23	19	15	22

Table 6. Number of correct items for participants 16-30 in Part B

	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25	P26	P27	P28	P29	P30
Q2	5	4	5	4	2	4	2	3	4	5	4	4	4	5	4
Q3	6	6	5	6	6	6	6	5	6	6	6	5	6	6	4
Q4	5	6	6	6	3	6	4	1	3	5	3	4	4	5	3
Q5	6	7	8	7	5	8	7	6	8	9	9	7	8	9	6
Total*	22	23	24	23	16	24	19	15	21	25	22	20	22	25	17

*Possible marks (Q2-5, Q3-6, Q4-7, Q5-10, Total=28)

As mentioned in the introduction to this report, Part B of the OET listening test involves various question types and requires test takers to listen to and extract information from a short lecture on a health related topic. In the shortened version of Part B used in the current study, question 2 was a sentence correction task in which test takers were required to cross out an incorrect word in the sentence and replace it with the word the speaker actually uses. Question 3 was a short answer task requiring test takers to extract specific details from the audio text. Question 4 was a sentence completion task in which some items required test takers to extract details from the audio text, while some items required test takers to paraphrase complex sections of the text in order to appropriately complete the sentence. Question 5 was a lecture note completion task, which required some limited paraphrasing as well as the extraction of specific details.

Each question is comprised of several items, and responses to each item are marked as correct if they correspond with answers listed in the marking guide. Tables 5 and 6, above, show the number of correct answers recorded by each participant for each of the four questions. As expected, questions 4 and 5, which required information to be synthesized and paraphrased, and as a result demanded a more in-depth level of text comprehension, were more difficult for most participants than questions 2 and 3, which required test takers to identify and write specific words (Q2) or specific details of more than one word (Q3). Overall, 7 participants answered between 11 and 15 items correctly out of a possible 28, 17 participants answered between 16 and 23 items correctly, and 6 participants answered 24 or more items correctly. Accordingly, verbal report findings, reported below, are summarised into three broad participant groups: (i) Participants with 11-15 correct items; (ii) Participants with 16-23 correct items; and (iii) Participants with 24 or more correct items.

Verbal reports

(i) Participants with 11-15 correct items

While most participants in the study appeared to have few, if any, difficulties with questions 2 and 3, participants in this group reported some problems in relation to both. They tended not to understand the text, but rather over-relied on a strategy of identifying key words in the items and listening for those words as a means of locating the required information. While this strategy was successful for most of questions 2 and 3, problems arose when the key word selected was misguided:

P2: *I just ah you know focus on the, the number come up so I pay more attention… a signal word yeah so when they have I put severe, severely but finally I found oh number six have severely so yeah… …*

Similarly, problems arose when a key word could not be readily identified:

P2: *Yeah I can't, I don't know which are, which are, which is the key words*

In the more difficult questions, 4 and 5, a lack of easily identifiable key words led to confusion as participants were not able to follow the audio or to identify which sections of the audio related to particular items:

P8: It's like I don't know where to concentrate, whether to look at the question or listen to what he has to say. It's like um I don't know where to focus should I focus on what he's trying to say or should I focus on the answer the question sorry not answer the question that they're asking

P9 was using key words from the items to cue and direct listening attention, and ran into difficulty when the words in the item were a paraphrase, rather than an exact match, of the audio text. The key words he was waiting for did not occur and as a consequence he missed the cue to move on and so missed information needed to answer subsequent items:

P9: I missed ah ah three question ah, because I didn't hear- ah ah without any treatment and I was waiting and he I and I missed this part … …I was waiting for about something and ah, and ah, I missed it, and ah I was ah ah waiting to ah say something about something and ah I missed mm and I heard that ah he ah talked about two more and I ah confused

P14 reported experiencing difficulty when it was necessary to understand the text rather than to simply identify particular words, particularly in relation to items which allowed for more open ended responses:

P14: Ah… like first question was still like the open question which is which is very hard for me to pick out the answers, I need to understand the whole things, yeah this part of the listening is very very hard yeah but uh, uh I tried to write down as much I can because I'm not sure which which which answer is the right one so I can write it down the whole things I write and I heard, so yeah

In question 5, most of the items required test takers to extract specific details from the audio, but the abundance of numbers in the text caused difficulty and confusion for participants in this group:

P23: Yeah I think this was a little … for me because all the information coming in ah, for me I can describe it as a big amounts of information and especially the numbers I didn't answer the number. I think I missed because he saying a lot of numbers… …

P5: It's fast um when he mentioned about the numbers I, I I I didn't understand like ah what he talking about so I, I I got lost so an-… and I couldn't write, write ah any question from here so… I think too fast…

Participants in this group also mentioned experiencing difficulty reading through the items in the time allocated:

P2: Umm, I think I didn't, I don't understand the question. So how many trials were they are… yeah ah um I read the, the question too slow, yep

Speech speed was also reported to cause difficulties for participants in this group:

P2: Just very fast, I don't have time to ah think and put in the right order.

P23: And the question was clear but my problem was with time, ah sorry, with the speed of his talking, his talking he spoke a little but fast.

<u>(ii) Participants with 16-23 correct items</u>

Few difficulties were reported by this group in relation to questions 2 and 3. As with the previous group, some participants here also reported using a strategy of identifying key words in the items as a means of focussing attention on the relevant parts of the audio, but for the most part without problem:

P6: this one particularly was very… very clear to understand and all the questions an-… I was I was concentrating on just one word was one word I was looking for one word and the clue should be... somewhere around in the middle of the sentence, basically… and yeah, was thinking of all the things that he said and... yeah… no no not so much for anything else

Most participants in this group, however, reported following and understanding each of the sentences. For example:

P12: Just I listen I uh look at the first uh… uh word in the sentence so when he uh when he start to starts to talk about it, I understand that the sentence is uh uh I mean that he is going to uh uh say this sentence, and I just follow his talking and I can find what's the wrong word

P13: Oh OK because uh he's speaking the same, exactly the same sentence so it's very easy to find which word is different, so I think this one is easier for me, I feel, yeah, and yeah, as long as he is speaking the same order the same sentence I feel it's ok for me to find

P19: I was just following word by word actually and I was just listening is the meaning here the same as that what he is saying.

Participants in this group were also able to identify key words in the audio rather than relying exclusively on finding key words in the test items, suggesting that they had understood the text and were able to target relevant details. For example:

P6: *and… in number five… ah… he mention two things about ah the medical condition so… the key words were about ah reduce and increase these was the main the key words… so… and this is cardiovascular disease… and the other one was breast cancer*

P17: *this one it says two medical conditions… for so I just pay attention to anything related to a disease, and he says breast cancer and cardiovascular disease so I just uh wrote them down like that… yeah*

Several participants mentioned having difficulties processing large amounts of information under time pressure in order to extract specific details for the last items in question three, as exemplified in the extracts below:

P3: *just I want to catch something relevant to the topic to write it but he still he discuss something before, before coming to the to each questions so I, I ah um maybe um maybe I will be confused between, to write this relevant or wait him for to wait him to speak another ah relevant ah information here, required here*

P13: *Uh, not quite sure, how many words should I write down, and the obviously I don't have enough time to write down the whole sentence so… mm… …yeah, because they say a lot of things for one question … so, how many things sh- they want us to write down*

Participants in this group also tended to engage comprehension monitoring as they followed the audio text and proceeded through the question, and so they were able to identify comprehension problems and make some corrections. For example, in relation to question 5, P3's comments reveal that he used existing world knowledge as a tool for identifying a problem with his initial understanding of a number mentioned in the audio text. '50' did not seem possible in the given context, so he was able to correctly infer that what he had actually heard was '15':

P3: *I could hear that he mentioned 50 OK but I I I didn't think it's 50 I think it's 15 OK 15 years it will be more it will be logic to be 15 not 50 just I just I ah um just I ah I thought it's as an age 50 so it's 15 or more years here it would be more logic, more logic*

As with the first group, speech speed was reported to cause increased difficulty for participants in this group:

P10: *He was too fast, especially here… don't know what he said*

P15: *I didn't follow the second sentence I was just going on it was a bit fast for me, I think it's risk less for the oestrogen therapy than the combined therapy, but I really didn't follow that because it was a bit quick for me*

<u>(i) Participants with 24 or more correct items</u>

As expected, participants in this group reported that they were able to follow the audio and answer items in questions 2 and 3 with no difficulty.

Of relevance to the test specifications for Part B, however, several participants in this group (and also the other two groups) answered the final item of question 2 incorrectly. The item reads as follows:

'And 20 per cent have symptoms which severely impact their quality of health'

Test takers were required to cross out 'health' and write 'life'. 13 out of the 30 participants wrote 'very severely' or 'very' for this item, perhaps because in the other items in question 2, the audio text matched the sentences exactly, apart from the wrong word, whereas in this item there was a slight difference - the word 'very', which appears before 'severely' in the audio, was omitted in the sentence. The modification was most likely made for formatting reasons, so that the complete sentence would fit on one line, but clearly this caused some confusion for participants, many appearing to assume that the missing word was the detail that needed to be corrected in the sentence. Participant 21, for example, answered this item incorrectly but reports no difficulty understanding the text:

P21: And over here he said very severely not just severely instead of severely. But this was really easy as compared to the previous one because the accent was really good and he was slow I could understand.

Even those who answered the item correctly reported some confusion with the item:

P1: I found there is little bit ah little bit um difference what he was saying and what it's what it was written here, 'cause it's very rich very severely we missed one word in this sentence is and I was going write that down but then I felt I need to cross some word and then I found this one is the wrong… so yeah I wrote this down

Particularly in relation to the more difficult items in questions 4 and 5, participants in this group were better able to perform comprehension monitoring and identify errors as they followed the audio and completed the items than participants in the second group. This was particularly evident in relation to one of the items in question 4, which many of the participants in the previous group answered incorrectly:

P4: *Um yes, the second question I, I didn't, I missed the word 'typically' so I grabbed the first age um what's that the age, the age group that he said but then um as he was talking more about the ah the ah the typical age I went back to the question and I saw the word typically written so I had to make the correction*

P21: *The second thing which was asked was particularly women who would be considered for hormone therapy are aged between 45 to 55 years. Before that he mentioned the age group of 50 to 79 years and I got confused, instead of skipping I actually wrote it here, then the other thing came in and I wrote it here because I could correlate and then I scratched the first because I thought that this was wrong.*

Furthermore, these participants were generally better able to handle lengthy and complex sections of audio, and to take and match cues from the audio and the item to successfully extract the relevant information. For example:

P1: *And the last one, is… um… it's a little bit harder as well but the last sentence he ah mentioned… they suggested like they… they think they should be cardiac protected but it's not so I got the answer from like his last word yeah*

P18: *he said it increases the risk of cardiovascular diseases so I thought so, at that point I just read the second result and hormone therapy um… so I said increase the risk of cardiovascular then when I, just looked back it said it was 'surprising', that word was like ok so it should be the opposite so I said it usually was cardio protective and it was surprising cause it increased the risk so yeah.*

In general, participants in this group were able to effectively select key word cues to direct listening attention from the items and from the audio text, and to move as required to a strategy of focussing on achieving an understanding of the gist or of particular sections of the text when focussing on key words proved inadequate. P18's comments, below, reveal this adaptability as she reports attempting to achieve an overall understanding to extract required information when the lecture notes do not correspond directly with the audio text:

P18: *it depends on, if the first two letters, ah-s first two words are what the audio says then I'm like OK I'll just follow the line… but um, if he just keeps talking in a general note, then I'm like OK let me just have the overall picture and see what's going on there…*

When participants in this group felt that they were unable to follow and understand parts of the audio text, as with participants in the other groups, they identified the demands of attending to listening and writing at the same time as a source of difficulty:

P1: *the increase risks similar to, this one is a little bit um… hard, 'cause you need to write more words than numbers so it's a little bit time consuming, so when I was writing it jumped to the next section and I was writing and I was hearing the next words.*

P25: *Um, difficulty in the last ah treatment type I became a bit overwhelmed with less or more because I couldn't finish reading up to this. I read up to here [increased risk].Yes, so um here I* *had to the same thing, I was reading and listening and that time actually he was saying lots of less, more, less, more so yeah I had a bit of problem here.*

Finally, and of further relevance to the test specifications, several participants reported experiencing confusion over where to direct listening attention and where to write answers due to the formatting of the first part of question 5. In the lecture notes, the words in bold ‘of 1000 women aged 50 to 70…’ occurring after the first item was intended as a topic heading to signal that information for the next items would follow. Many of the participants misunderstood and thought the topic heading, above, was the stem for the second item and that they needed to write a response to replace the ‘…’:

P4: *So um, I saw these dots here, so I thought I have to fill in what he’s talking but it was actually supposed to be filled in the blanks under… so I think I kind of got the… was able to go with him and fill the rest of the blanks yeah*

P21: *Yeah, um this was pretty straight forward he actually said 1997 the first time and the other thing was I thought that this dot, dot, dot means there is some gap, but this wasn’t something to be filled up*

Summary of results for Part B

As mentioned in the introduction of this report, the abilities listed for Part B in the test specifications overlap with those listed for Part A. The different question types of Part B are intended to achieve a broader and more nuanced spread of difficulty across the test by allowing specific abilities to be targeted directly and somewhat in isolation from other abilities. The specifications for Part B distinguish ‘direct meaning comprehension’ from ‘inferred meaning comprehension’ as follows:

- Direct meaning comprehension
    - Understanding an overarching argument or point
    - Understanding main ideas and important information
    - Listening for specific details
    - Understanding health-specific vocabulary

- Inferred meaning comprehension
    - Making inferences from information available in the text
    - Inferring meaning of unfamiliar lexical items from context
    - Recognising the communicative functions of utterances, according to the context of talk, the speaker and his/her goals

Again, the shortened version of Part B used in the current study consisted of four question types: Q2 – sentence correction, Q3 – short answer, Q4 – sentence completion, Q5 – lecture note completion. While some of the weaker participants experienced difficulties with questions 2 and 3, most of the participants in the second and third groups had few if any problems with most of the items. The difficulties that arose for the weaker participants were due to a lack of direct meaning comprehension, and an overuse the strategy of relying on key words in the items to direct listening attention. Even so, one of the participants in the first group, P14, was able to answer all items for question 2 correctly, reportedly without understanding the meaning of the text or the item sentences. There is, therefore, evidence that question-types such as sentence corrections (Q2) tap into 'listening for specific details', listed under direct meaning comprehension, without necessarily drawing on other abilities listed there. This supports the appropriateness of this question format for measuring this particular ability, as listed in the specifications, and to identify words and word boundaries, but the question-type clearly does not necessarily tap into the more complex listening processes involved in text comprehension. As such, items in this format are useful for distinguishing between weaker participants and less useful for distinguishing between test takers at higher levels.

Question 3 caused problems for a few participants in the first two groups, firstly because a strategy of identifying key words was not always successful, and weaker participants in particular tended to over-rely on this particular strategy, as mentioned above. Secondly, while the items in this question asked for specific details from the audio text, the more open ended nature of the final items meant that participants were required to follow and understand lengthy sections of text to locate and extract the answer. Consequently, evidence support the use of this question type to elicit both a narrow range of abilities, as some participants were able to identify required answers to the first items with a very limited understanding of the text, as well as a broader range of abilities. Verbal reporting of difficulties encountered when comprehension broke down in relation to this question support the appropriateness of this question type (short answer) as a means of eliciting an ability to understand the overarching point as well as main ideas and important information, for example.

Questions four and five of Part B were more difficult, with most participants finding question 4 the most difficult overall (in terms of their verbal reports as well as the number of correct items). The final item of question 4 involved the need to understand and summarize complex sections of audio text in order to provide an accurate and appropriate sentence ending, which few of the participants were able to do with ease. Verbal report data showed evidence that some items in these questions tapped into aspects of inferred meaning comprehension. The following extract, also cited in the results above, illustrate, for example, that P18 is making inferences from information available in the text, as listed in the specifications:

P18: *he said it increases the risk of cardiovascular diseases so I thought so, at that point I just read the second result and hormone therapy um… so I said increase the risk of cardiovascular then when I, just looked back it said it was 'surprising', that word was like ok so it should be the opposite so I said it usually was cardio protective and it was surprising cause it increased the risk so yeah.*

Further, earlier items in question 4 required specific details to be extracted, but necessarily drew on more complex abilities listed as aspects of direct meaning comprehension, such as understanding the overall point as well as main ideas and important information, as did the more difficult items at the end of question 3. Question 5 required participants to identify and select specific details from the text, and most participants reported experiencing difficulty due to the speed of text and the density of the information, especially as information needed to be paraphrased. At times, participants were able to paraphrase the meaning of the audio text in their verbal reports, but were unable to synthesize the information appropriately in writing to complete the sentence as required.

As was discussed in relation to Part A, the question of whether or not written summary skills impact performance on the listening test demands further attention. Further, and not surprisingly, some participants reported difficulty due to the need to listen and write simultaneously in Part B, as they did with Part A, here especially when some of the items were long and could not be properly read during reading time. Similarly, the length and complexity of item stems in Part B of the test might be reviewed in order to verify that reading ability is not a significant construct irrelevant variable.

Discussion and conclusion

The primary aim of the current study was to explore the construct validity of the OET listening subtest by investigating if the processes test-takers reported engaging in resembled those which the task is designed to elicit. Data collected using verbal reports provide evidence in support of the validity of the task, as many of the listening processes and strategies reported by test-takers mirrored theoretical expectations, and called upon abilities listed in the task specifications. Evidence suggests that while the two parts of the test, A and B, draw on similar types of underlying phonological, syntactic and semantic knowledge, the note taking task in Part A and the various question types comprising Part B made different demands on test takers. Consistent with previous findings reported by Vandergrift (2003), the aspects of linguistic knowledge as well as other non-linguistic knowledge resources, such as general and professional knowledge, that participants needed to draw upon varied to some extent based on task-type.

Not surprisingly, participants reported difficulties in relation to both parts of the test due to unfamiliar vocabulary. In the note-taking task (Part A), however, this was particularly problematic for weaker participants (those who answered the least number of items correctly), who were often unable to distinguish word boundaries and identify unknown words. This led to them missing specific details and often failing to understand the gist of sections of the audio text. Those who recorded the highest proportion of correct items, by contrast, were often able to identify and accurately record unfamiliar words, even if they were unsure of the meaning. Better performing participants were also more likely to use contextual clues and existing linguistic knowledge to guess at the likely meaning of novel lexical items. They were also able to drawn on professional and world knowledge, in combination with existing linguistic knowledge, to make accurate inferences about what the speaker intended to say when sentences or words were truncated or unclearly articulated.

An ability to extract relevant information is defined in the test specifications as a key aspect of a candidate's ability to "follow the facts" in a consultation between a health professional and a patient, and consequently this was a focus point in the analysis of verbal report data for Part A. Song (2012) argues that while note taking tasks are a good indicator of listening proficiency as they allow test takers to demonstrate whatever they understand, test takers "might be inclined to take notes of whatever they want to, including details, even if they do not clearly understand the interconnectedness among the ideas" (2012: 83). In the current study, weaker participants did tend to write down all details they could understand and identify, consistent with Song's claim. Further, participants were able to identify and record correct details without understanding the gist of the audio text in some instances.

Relevant to Part A and also to the more open ended questions in Part B of the test, Buck (1991) notes that if the amount of information required in the response is not made clear in the question, then test takers can be unsure of how much information to include. He found that test takers tend to write all that they hear and as a result, run out of time and fail to answer subsequent questions. This is also partly consistent with findings in the current study, in that several participants reported missing subsequent information because of the amount for detail they felt they needed to include for some items.

In general, however, for Part A most participants understood the need to select information for relevance in this task, and to write main points in note-form. Several participants reported drawing on key words in the topic headings, as well as existing professional knowledge in order to select relevant information. It should be noted, though, that despite understanding the task requirements, many participants reported confusion and difficulties concerning how much they needed to write, and at times about whether or not particular information should be considered relevant. Most of the items in Part B were not open ended, and so item stems directed test takers in terms of identifying relevance, but where items required information to be synthesized and summarised, only the better participants were able to respond correctly.

As stated previously, in relation to Part B of the test, few problems were reported in relation to questions 2 and 3 as participants were able to rely on a strategy of identifying key words to direct their listening attention, and often could identify the required information without needing to understand the meaning of the text. As already noted, this supports the appropriateness of this question format for measuring test takers' ability to listen for specific details, as listed in the specifications, and to identify words and word boundaries, but clearly does not tap into the more complex listening processes involved in overall text comprehension.

Of further relevance to the test specifications for Part B, as noted earlier in the results section, participants experienced confusion when one of the items in question 2 consisted of a sentence that, in addition to the word that needed to be crossed and replaced, was slightly different from the audio text. This item probably caused confusion because all of the other items had matched the audio word-for-word except for the single error that needed correction. While the instructions for this question type are clearly stated, in order to avoid confusion the test specifications should possibly be amended to specify that the sentences must match the audio exactly, or the question instructions should be amended to make it clear that the sentences may be paraphrases of the audio text.

Questions four and five of Part B appeared to be more difficult than questions 2 and 3, with most participants finding question 4 the most difficult overall (in terms of their verbal reports as well as the number of correct items). As noted in the summary of results for Part B, this question, as well as some items in questions 3 and 5, tapped into more complex aspects of direct meaning comprehension, and also inferred meaning comprehension. A formatting issue was also identified as a source of confusion in question 5, suggesting the specifications need clarification.

As expected, findings from the current study indicate that the cause of comprehension difficulties vary from participant to participant due to their particular gaps in linguistic knowledge and non-linguistic resources. On the whole, however, results show clear differences between participants according to the number of items they answered correctly. For both Parts A and B of the listening subtest, participants who achieved the least number of correct items generally displayed a lack of word recognition and an inability to identify and reconstruct the main ideas from the audio text.
Buck (2001) highlights the distinction between controlled and automatic processing as particularly relevant in second language listening, and findings here suggest that participants with the most correct items had greater access to automatic processing than other participants, which allowed them to better meet the cognitive demands of listening and writing simultaneously, and of needing to process and summarize complex information as well as to make inferences as required. Further, the various question types in Part B allowed specific abilities, particularly lower level abilities, to be effectively targeted, hence allowing the test to capture differences in listening abilities among weaker participants as well as participants with higher level listening comprehension skills.

Verbal report data from stronger participants suggest that they engaged in verifying their interpretation and understanding of the texts as they progressed through the tasks in Part A and Part B, making corrections and adaptations as needed, whereas this was absent from reports provided by weaker participants. Vandergrift (1997, 2003) refers to this as the metacognitive strategy of comprehension monitoring. Vandergrift (2003), in his comparison of strategy use by skilled and less skilled second language listeners, also found that more skilled listeners used metacognitive strategies, primarily comprehension monitoring, more frequently and effectively, and were able to interact more deeply with the text than their less skilled counterparts. Findings in the current study also suggest that better participants were able to construct a more complete "textbase" and "situation model" as they listened (Van Dijk and Kintsch, 1983), allowing them to check for consistency between their mental representations and information from subsequent propositions in the text on an ongoing basis.

Finally, in relation to comprehension monitoring, Buck claims that despite its importance in listening comprehension, particularly second language listening comprehension "where linguistic knowledge and processing efficiency may be grossly inadequate and the listener is often trying to interpret a text from a partial analysis of the propositional content" (1991: 80), it is unclear if and how this could be tested. Findings from the current study provide evidence to suggest that some texts and item types may combine to elicit this strategy. For example, complex and information dense sections of audio text, particularly when combined with open-ended item types requiring test takers to understand the text in detail in order to extract relevant information, as well as to then summarise the information in an appropriate form (such as sentence completions – Part B, question 4) appeared to prompt this sort of strategy from participants.

In conclusion, as stated at the outset of this report, the aim of the current project was to use verbal reports to gather construct-related evidence in order to evaluate the explanation inference, one of the six principal inferences underlying the interpretation of test scores in an argument-based approach to test validation (Chapelle, Enright and Jamieson, 2010; Kane, 1992; Kane, Crooks & Cohen, 1999; Xi, 2008). The explanation inference, which relies on the assumption that the listening processes, skills and strategies elicited by the tasks are captured by the test specification, was, on the whole, supported by this investigation. Verbal report data provided strong evidence that the two parts of the OET listening sub-test tapped into key abilities listed in the test specifications, and that the different parts of the test made different and appropriate demands on test takers. The data also provide evidence that the different task types in Part B create varying degrees of difficulty in terms of the range of strategies and abilities participants were required to utilise in order to complete each question. This study thereby provides valuable evidence in support of the overall validation argument for the OET.

RECOMMENDATIONS

On a practical level, an aim of the study was also to gain insights into test taker listening processes and strategies in order to inform refinements to the listening test specifications and to thereby enhance future test design. Based on these findings, the development of more detailed task specifications is recommended.

In particular, it is recommended that the specifications are clarified in relation to formatting guidelines and instructions to test takers where problems have been identified herein. This study also allows for more specific instructions for item writers in terms of achieving a broad and measured spread of item difficulty, through insights obtained into the relationship between text features, item-type and difficulty. Such a refinement of the test specifications should ensure that different listening task-types can be more specifically matched to particular strategies and aspects of listening abilities, and should also ensure that a measured and consistent range of item difficulty is included in each task version.

Finally, potential sources of construct-irrelevant variance brought to light in the study, such as the impact of written summary skills, reading ability and spelling knowledge, might be investigated further in order to rule out any significant influence on test outcomes.

References

Berne, J.E. (2008). Listening comprehension strategies: A review of the literature. *Foreign Language Annals*, 37 (4), 521-531.

Bourne, L.E., Dominowski, R.L., & Loftus. E.F. (1979). *Cognitive processes*. Englewood Cliffs, USA: Prentice Hall.

Buck, G. (1991). The testing of language listening comprehension: an introspective study1. *Language Testing*, 8 (1), 67-91.

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

Chappelle, C., Enright, M. & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29, 3-13.

Conrad, L. (1981). Listening comprehension strategies in native and second language. (Doctoral Dissertation, Michigan State University, 1981). *Dissertation Abstracts International,* 42, 690A.

Conrad, L. (1985). Semantic versus syntactic clues in listening comprehension. *Studies in Second Language Acquisition*, 7 (1), 59-72.

Flowerdew, J. & Miller, L. (2005). *Second language listening: Theory and Practice*. Cambridge, UK: Cambridge University Press.

Gass, S. & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Goh, C. (1998). How ESL learners with different listening abilities use comprehension strategies and tactics. *Language Teaching Research*, 2 (2), 124-147.

Goh, C. (2000). A cognitive perspective on language learners’ listening comprehension problems. *System*, 28, 55-75.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge, UK: Cambridge University Press.

Gruba, P. (1999). *The role of digital video media in second language listening comprehension.* Unpublished doctoral dissertation, University of Melbourne, Australia.

Harley, B. (2000). Listening strategies in ESL: Do age and L1 make a difference? TESOL Quarterly, 34 (4), 769-776.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.

Kane, M., Crooks, T., and Cohen, A.(1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18, 5–17.*

Lumley, T. & Brown, A. (2005). Research methods in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 833-856). Mahwah, New Jersey: Lawrence Erlbaum Associates.

McNamara, T. (1990). *Assessing the second language proficiency of health professionals*. Unpublished doctoral dissertation, The University of Melbourne, Melbourne, Australia.

Martin, T. (1982). *Introspection and listening process*. Unpublished master's thesis, University of California, Los Angeles, USA.

Nagle, S.J., and Sanders, S.L. (1986). Comprehension theory and second language pedagogy. *TESOL Quarterly*, 20 (1), 9-26.

Ockey, G. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, 24 (4), 517-537.

O’Malley, J., Chamot, A., & Kupper, L. (1989). Listening comprehension strategies in second language acquisition. *Applied Linguistics*, 10 (4), 418–37.

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

Richards, J.C. (1983). Listening comprehension: approach, design, procedure. *TESOL Quarterly*, 17(2), 219-240.

Song, M. (2012). Note-taking quality and performance on an L2 academic listening test. Language Testing, 29 (1), 67-89.

Van Dijk, T.A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. London: Academic Press.

Vandergrift, L. (1997). The strategies of second language (French) listeners. *Foreign Language Annals*, 30 (3), 387-409.

Vandergrift, L. (2003). Orchestrating strategy use: Toward a model of the skilled second language listener. *Language Learning*, 53 (3), 463-496.

Wagner, E. (2008). Video listening tests: What are they measuring? Language Assessment Quarterly, 5 (3), 218-243.

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.

Xi, X. (2008). Methods of Test Validation. In E. Shohamy and N.H. Hornberger (eds). *Encyclopedia of Language and Education*, 2nd edition, Volume 7: Language Testing and Assessment, pp. 177-196.

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.