1 The cognitive validity of the lecture-based question in the IELTS Listening paper

Author

John Field
University of Reading, UK

Grant awarded Round 11, 2005

This study investigates the cognitive validity of one part of the IELTS Listening Test by comparing the performance and experience of subjects when completing a sample IELTS task with their performance and experience when doing a similar task that is not constrained by the conditions of test administration.

ABSTRACT

This study investigates the cognitive validity of two samples of IELTS lecture-listening material taken from past papers. In one condition, participants were asked to undertake the original test, and to provide a retrospective verbal report, explaining why they had chosen the answers that they had favoured. In a second condition, they were asked to take notes under the less constrained circumstances that obtain during a lecture, and then to report on them. The material was distributed on an ABAB principle so that the possible effects of recording and test method could be compared.

The scores obtained by individuals under 'test' conditions were compared with the extent to which the same individuals showed themselves capable of accurately reporting units of information in the freer 'lecture' condition. No clear correlation was demonstrated.

The verbal reports were then examined for evidence of the cognitive processes adopted by participants under test conditions, and were matched against conventional psycholinguistic accounts of first-language listening (see, for example, Brown 1995, Field 2009). A distinction was made between *normal processes* which might equally well be adopted by a native academic listener, *strategic behaviour* which aimed to compensate for problems of understanding; and *test-specific behaviour* representing the user's response to characteristics of the test. Evidence of the last raised concerns about cognitive validity. The protocols showed participants adopting specific routines that were tailored to the test method. They also provided considerable evidence of participants favouring test-wise strategies and attempting to exploit loopholes in the format of the test such as the availability of questions in a written form.

A third line of enquiry investigated participants' responses to listening under the two conditions, in order to establish which they had found the more demanding. An unexpected result was the number of participants who found lecture listening less demanding than undertaking the test. Possible reasons are explored.

AUTHOR BIODATA

JOHN FIELD

Dr John Field's research interests lie in first and second language listening and in the application of psycholinguistic theory to an understanding of second language processes. He teaches psycholinguistics and child language development at the University of Reading and cognitive approaches to SLA at Cambridge University. He has guest edited issues of international journals on both L2 listening and psycholinguistics in ELT, and has written background books on psycholinguistics, among them a widely-used reference source. He has advised on the cognitive validation of writing, reading and speaking papers in the Cambridge suite. His Listening in the Language Classroom (CUP) appeared recently.

IELTS™
English for International Opportunity

The test that sets the standard

IELTS RESEARCH REPORTS
VOLUME 9, 2009

Published by: British Council and IELTS Australia
Project Managers: Jenny Holliday, British Council Jenny Osborne, IELTS Australia
Acknowledgements: Dr Lynda Taylor, University of Cambridge ESOL Examinations
Editor: Dr Paul Thompson, University of Reading, UK

ISBN 978-1-906438-51-7 © British Council 2009 Design Department/X299
The United Kingdom's international organisation for cultural relations and educational opportunities.
A registered charity: 209131 (England and Wales) SC037733 (Scotland)

CONTENTS

1 BACKGROUND

1.1 Need for the study

Cambridge ESOL takes pains, in designing the specifications of the IELTS Listening paper, to ensure that the test achieves content validity (cf Clapham 1996, pp 65-72, on content validation for the parallel reading paper). The criteria ensure that the four listening passages are closely linked to an academic context by specifying the situations and text genres that candidates are likely to encounter, either socially or in the course of study. In this way, the designers ensure that the test achieves validity in relation to linguistic factors such as the lexical, discoursal and pragmatic content of the target field.

However, IELTS is first and foremost a test of language *skills*. It serves as a predictor of performance, on the assumption that its results correlate with a candidate's ability to handle the real-world demands of an academic programme. It thus has to be evaluated in terms of a second type of construct validity, namely cognitive validity (Glaser 1991; Weir 2005). In Weir's (2005) evidence-based validity framework, the term refers to the extent to which the cognitive demands of the test reflect those of the target context. In relation to the IELTS Listening paper, this entails establishing that the types of listening behaviour which the test elicits correspond to those which an academic environment requires.

Traditionally, this type of validation is conducted in a post hoc fashion, with statistical methods such as factor analysis applied to test results in order to establish the nature of the construct which has been tested. Weir expresses concerns over this approach, raising the issue of whether the data under examination might not to some extent be compromised by the form and content of the test and by the assumptions underlying its design. In effect, he draws attention to the dangers of relying exclusively on an approach that attempts to track back from a product to the process that gave rise to it. He argues instead for what he terms *theory-based validity* (or, more recently, *cognitive validity*): a complementary approach to test validation which takes account, before the test is designed, of external empirical evidence concerning the nature of the construct that is to be assessed. Weir makes his point powerfully:

> There is a need for validation at the *a priori* stage of test development. The more fully we are able to describe the construct we are attempting to measure at the *a priori* stage, the more meaningful might be the statistical procedures contributing to construct validation that can subsequently be applied to the results of the test. Statistical data do not in themselves generate conceptual labels. We can never escape from the need to define what is being measured, just as we are obliged to investigate how adequate a test is in operation. (Weir 2005, p 18)

This additional strand of construct validation requires that, alongside benefiting from feedback from piloting and past administrations, test design also draws in a principled way upon external evidence concerning the nature of the expertise which is to be targeted.

As noted, insights into the processes applied by candidates are especially important in the case of tests which are used to predict later performance. It is precisely these predictive tests which are worst served by a product-based approach. A researcher might indeed employ factor analysis to indicate the aspects of the target construct that have been tested; or might compare the rankings achieved in the test to other measures of the candidates' current competence. But neither result demonstrates the candidate's ability to perform on arrival in the target setting. The obvious way such a finding can be achieved is longitudinally – by measuring achievement once the academic programme has begun – but here the researcher faces a potential confound. If one uses overall measures of achievement during the course of the programme, it becomes difficult to separate the candidate's flair for the chosen subject of study from his/her L2 study skills. Similarly, one can trace broad correlations between overall IELTS scores and overall academic success; but it is difficult to single out performance in specific skills areas.

One possible line of enquiry for cognitive validation is to seek evidence in other disciplines – in the case of language skills, from the detailed models of L1 processes which cognitive psychology has been able to build on the strength of long-term empirical findings (see Shaw and Weir 2007, and Field, forthcoming, for applications of this approach to the cognitive validation of L2 skills tests). An alternative course is to seek evidence not *a priori* as Weir proposes, but on-line, while a task is being carried out. Comparisons can be made between the observed or reported behaviour of the candidate when performing the task under test conditions and the candidate's parallel behaviour when the task is performed under conditions which more closely resemble those of the real life context for which he/she is preparing. Such evidence meets Weir's strictures in that it is not tied narrowly to test outcomes but directs enquiry to the processes which give rise to those outcomes.

The present study adopts the second approach. It investigates the cognitive validity of the IELTS Listening paper by comparing the performance of participants during sample IELTS tasks with their performance during a task which more closely replicates the demands of an actual academic context. By 'performance' is to be understood both evidence of successful comprehension and evidence of the processes that are employed by the candidate to achieve that goal. The process evidence will be considered in relation to established models of listening drawn from cognitive psychology and underpinned by extensive empirical evidence. In this way, Weir's plea for the greater use of external, scientifically validated information will be met.

The IELTS Listening paper falls into four sections. The first two contain recorded content that relates to what are termed 'social needs'; the third tests ability to understand a conversation with details of course content and assignments; the fourth tests lecture-listening. Of these, it is the last which affords the most compelling case for cognitive validation. Its predictive validity rests heavily upon the extent to which it can be shown to model performance in a one-way lecture-listening situation (admittedly, with some limitations such as lack of visual support). It is thus from this section that the present study selects its material.

1.2 Cognitive validity and L2 listening

Weir (2005) stresses the importance of applying criteria that are based upon an understanding of the processes underlying the L2 skill to be tested. However, he leaves the precise nature of those criteria to some extent open to discussion. Reflection suggests that one might establish the benchmark for cognitive validation in two different ways:

a by treating predictive validity as a primary criterion; and comparing the processes in which non-native listeners (NNLs) engage when performing a particular task under test conditions with those which they employ under non-test conditions.

b by treating native-like performance as a primary criterion; and comparing the processes in which NNLs engage when performing a particular task under test conditions with the processes adopted by native listeners (NLs) in real-life conditions.

Given the important predictive role of IELTS, the focus of the present study is upon the first.

When one considers the question of cognitive validity with specific reference to L2 skills (and within them L2 academic skills), it is important to differentiate between three different types of behaviour. They are referred to generally as 'processes'; but a distinction needs to be made between:

behaviour which is part of the *normal process* – in the present case, behaviour which might equally well be adopted by a native academic listener

strategic behaviour which aims to prepare for a task, to maximise the amount that is retained or to compensate for problems of understanding. In listening, much of this behaviour will be specific to the L2 listener in that it anticipates or deals with problems of understanding that are due to individual perceptual or linguistic limitations. (Note that the term strategy is used rather more narrowly than, for example, in Buck 2001, pp 103-4)

test-specific behaviour representing the user's response to features of the test. It would seem to take two distinct forms. The candidate might adopt specific routines which assist in the achievement of the particular task set in the test, but which would not normally play a part in the corresponding real-life activity (in the present case, lecture listening). Or the candidate might adopt certain testwise strategies in an attempt to second-guess the intentions of the setter or to exploit loopholes in the format of the test such as (in listening) the availability of questions in a written form. Clearly, either of these constitutes a negative factor when attempting to establish cognitive validity.

The present study aims to keep these three performance components as separate as possible. One needs:

a to seek parallels between the language processes involved when taking the test and those involved in listening to the same material when unconstrained by test conditions

b to seek parallels between the compensatory strategies applied to problematic areas of the input when a participant is under test and under non-test conditions

c to identify strategies specifically related to test-taking, which raise possible concerns about cognitive validity.

1.3 Choice of method

The most appropriate method for the study was verbal report (Ericsson and Simon 1993). It has a number of disadvantages, which are acknowledged below. However, it is widely employed as a means of investigating various forms of expertise (including mathematical thinking and chess playing) and of identifying the operations which underlie them. Clearly, there are differences between the type of cognitive process which can be elaborated heuristically in terms of a set of consciously formulated stages and the type which entails a much less structured process such as deriving meaning from a text. However, both types of performance might be characterised as goal-oriented, and in both cases the goal (here in the form of the listener's answers) can be used as a means of tracking back to the thinking which gave rise to it.

Verbal report has been used successfully to investigate the processes of second language learners, who have proved capable of recording the thought processes which led them to particular interpretations of texts (Faerch and Kasper 1987). It has even been used (Cohen 1998) to research speech production and reception. Clearly, in the latter case, report has to be retrospective – which means that it is important to avoid memory effects. In fact, the circumstances of a listening test support retrospection well in that the participant has to provide a set of answers, which provide triggers to assist recall of the thought processes that led to them. In non-test conditions, the participant can be asked to write concurrent notes, which similarly support recall.

An important constraint of verbal report as a method should be mentioned at this point. Gathering and transcribing protocols is costly in terms of time, and consequently imposes limitations upon the size of the population that can be studied. Whereas it is possible to administer a test such as the IELTS listening paper over a very large population for the purposes of, for example, post-hoc factor analysis, a study that investigates individual on-line processing must inevitably draw upon a smaller group of respondents. The present project should be regarded in much the same way as a case study, though it reports on a larger number of respondents than do most. The numerical and statistical results recorded here must be regarded as broadly indicative rather than conclusive. That said, what is lost in generalisability will, it is hoped, be compensated for in the depth of the information that is obtained.

1.4 Theoretical framework

The present study bases its analysis upon a data-driven approach in which the researcher seeks patterns of similarity and difference in the responses recorded by participants with no *a priori* assumptions. However, any study of this kind also ideally requires a wider framework against which its findings can be measured.

Two possible theoretical areas suggest themselves within the literature on second language listening; but neither is extensive enough or well enough supported by rigorous empirical research. Firstly, there have been a number of proposals for taxonomies of listening sub-skills, of which the most notable are perhaps Richards (1983), Rost (1990, pp 150-158), Buck (2001, pp 57-59) and (specifically related to assessment) Dunkel, Henning and Chaudron (1993). But all of them contain categories with a degree of overlap, a lack of supporting research evidence based on listening in a natural context and no criteria to mark out certain characteristics as carrying more weight than others. A second possible source is the considerable work that has taken place in recent years on L2 listening strategies. It suffers from a number of theoretical problems – not least, the rather miscellaneous taxonomy adopted by many researchers and based upon Oxford (1990). As Alderson (2000, p 309) commented in relation to L2 reading, 'Much of the research into, and teaching of, reading strategies remains fairly crude... and frequently fails to distinguish between strategies as defined more generally in the strategy literature and "skills" as often used in the reading literature.' Much of the research (see e.g. Vandergrift 2005; Vandergrift et al 2006) has been dependent upon the use of questionnaires – a method which can at best only provide information about the strategies that learners *believe* they employ and is very much open to challenge in that it invites learners to provide information on processes that may not be accessible to report. Most importantly, an approach based solely on strategy use provides useful insights into the techniques employed by the listener in order to resolve local problems of understanding, but does not capture what is of equal concern in a study of cognitive validity, namely, the processes which a listener employs in decoding input and analysing meaning under circumstances that are unproblematic.

A more reliable theoretical framework is therefore found in the models of listening and of meaning construction which have been developed by psycholinguists investigating first language speech processing (see, for example, in Gaskell 2007, papers by Pisoni and Levi, McQueen, van Gompl and Pickering, Tannenhaus, Singer). They are elaborated in considerable detail, soundly based upon current thinking in cognitive psychology and underpinned by solid research findings. Granted, these are accounts of L1, not L2, language processing, but one can argue that, in identifying the traits of the skilled L1 listener, they provide a yardstick for assessing the performance of

the L2 listener at any level, and a goal towards which the EAP listener in particular might be expected to strive. Reference will be made to cognitive models of this type during the discussion. Particularly germane will be the ways in which they represent the cognitive demands that a given task places upon a language user.

1.5 Research questions

The present study investigates the extent to which the fourth section of the IELTS Listening text achieves cognitive validity

- by replicating the processes in which candidates would engage when listening to a lecture in a non-test content
- by measuring the ability of candidates to engage in the processes entailed in listening to a lecture in a non-test content.

The specific research questions are as follows:

1. To what extent can the fourth section of the IELTS Listening text be said to achieve construct validity in terms of the cognitive processes which it requires of the candidate?
2. How great is the role played in the fourth section of the IELTS Listening text by processes which are specific to the text context?
3. What are the perceptions of candidates as to the demands of the test when compared with those of listening to an academic lecture in non-test conditions?

The study first compares results achieved by means of the test with those achieved in a less constrained lecture-listening situation. Using verbal report, it then seeks evidence of the extent to which candidates taking the section 4 test employ test-wise strategies and other techniques specific to the test-taking context. It also examines the comments of participants on whether the test situation adds to or reduces the difficulty of academic listening.

2 RESEARCH DESIGN

2.1 General outline

A group of participants (N=29) were studied, all of whom were preparing for university entrance.

There were two conditions: Test and Non-Test. The Test condition entailed listening once to a passage from Part 4 of an IELTS Listening paper and supplying the answers required by the test setters. The Non-Test condition entailed listening to a Part 4 recording from another IELTS paper, making notes during listening and writing a short summary.

Validity would be compromised if participants were to hear the same listening passage twice. Two passages were therefore employed, and an ABAB design was adopted. Fifteen participants reported on Passage A in the Test condition and Passage B in the Non-Test; and the remaining fourteen reported on Passage B in the Test condition and Passage A in the Non-Test. So far as possible, each AB participant was paired with a BA one who shared the same first language.

After each task, participants were invited to describe

- the processes involved in achieving answers under test conditions
- the processes involved in extracting information and building meaning under non-test conditions

2.2 Methods

2.2.1 Verbal report

Verbal report is widely used in research into expertise generally (Ericsson and Simon 1993) and into cognitive validity specifically (Baxter and Glaser 1998). It has a number of drawbacks as a method of researching language skills performance (see McDonough and McDonough 1997, pp 191-200; Brown and Rodgers 2002, pp 53-78) especially in relation to the receptive skills and to non-native participants. They include the following:

a Thinking does not proceed on a step-by-step basis as it might in the resolution of a problem in (e.g.) mathematics or chess playing that involves logic.

b The reading and listening skills can only be investigated indirectly; and some of the processes involved may not be readily accessible to report.

c The process of reporting can interfere with the ecological validity of the task. In the case of listening, it is clearly impossible for participants to engage in concurrent verbal report. The use of retrospective report, however, carries possible memory effects.

d Language limitations may prevent non-native participants from reporting as fully as they might.

e The level of reporting may vary considerably from one participant to another – with implications for reliability.

One way of overcoming the memory effects associated with retrospective report is to provide 'stimulated recall' in the form, for example, of a video replay of the activity to be reported on (Gass and Mackey 2000). The importance of retrieval cues is well attested in memory research findings within cognitive psychology (for a non-specialist review, see Kellogg 1995, Chap. 5). Tulving's influential *encoding specificity hypothesis* (Tulving 1983) states that accurate recall is critically dependent upon activating the same cues in retrieval as those originally encoded with the event to be recalled. In the Test condition, such a trigger was available in the answers chosen by the participant. In an interview setting, the participant was asked to report his/her answers and then to explain the process by which the answer had been derived. In the Non-Test condition, the content of the participant's notes and written summary served similarly to provide a set of retrieval cues.

The approach adopted also attempted to reduce possible memory effects by ensuring a minimal time lapse between the process to be reported and the report itself (Brown and Rodgers 2002, p 55). The target listening passages were divided into three, providing pauses in which the test taker could record from 3 to 4 answers and report his/her thought processes after a relatively short listening period. The aim was to ensure greater detail and greater accuracy. The practice did not materially change the conditions under which the IELTS Listening Test is undertaken, since takers are allowed only one listening and thus have to record their answers in an on-line fashion.

Pausing the recording at appropriate intervals where there was a change of sub-topic was felt to be more ecologically valid than pausing it as each answer was achieved. The latter procedure would have been disruptive of the process of meaning building at a global level. It would also have meant that the researcher would need to signal the point at which the answer was identified, thus eliminating the uncertainty about matching a question to a possible answer that is an important feature of the experience of taking an L2 listening test.

So far as the non-test condition was concerned, the recording was paused only once and briefly, to ensure that the participant did not feel too challenged by the demands of note-taking.

Clearly in any research into listening and speaking, the verbal reports obtained need to be retrospective. Here, they were of two kinds.

a *In the Test condition*, participants reported each answer they chose and then explained their reasons for choosing it.

b *In the Non-Test condition*, participants were interviewed after writing a summary of the passage; and asked to report as much as possible of what they had heard in the recording (assisted by the notes they had taken and the summary they had written). They were allowed to decide for themselves the relative importance of what they reported and the discourse-level relationships between the different points.

A particular concern was that limitations of linguistic ability might be an obstacle to informative reporting. The two tasks, together with the reporting phase, were therefore piloted with six participants who shared the same background as the target group but had slightly lower overall IELTS scores. They provided to be capable of reporting clearly and accurately their reasons for choosing particular answers and ignoring others. Their comments also provided indications of the types of strategic decision that they had made.

2.2.2 Note-taking and summary

The original research design included a second source of data in the form of a written summary of what participants had heard in the non-test condition. Participants were to take notes while listening to the mini-lecture, and were then to write them up as comprehensively as possible. The purpose was to achieve hard evidence of how accurately and extensively each individual was able to report the mini-lecture on the basis of their notes. This would enable experienced judges to rate the participant's lecture-listening skills. The study would then seek possible correlations between the summary rating and the marks obtained in the IELTS test format.

Summary is a very informative method of testing listening comprehension skills, though it obviously poses practical difficulties of reliability and ease of marking in international tests (Alderson 2000, pp 232-3). Unlike more formal test methods, it provides evidence of ability to identify main points and speaker's purpose, to assess the relative importance of information and to show propositional links. It also requires the summariser to draw entirely upon information supplied by him/herself rather than using test items as a basis. Finally, it has some ecological validity in relation to a lecture-listening task, since clearly the content of some real-life lectures may ultimately find its way into a student's assignment.

However, the piloting phase raised questions about the value of using summaries in this particular project. Participants were told that they could take as long as they liked to write their summary, but in practice they often wrote very little. Two factors seemed to constrain them. The first was the face-to-face situation: they seemed to feel that their inevitably slow writing as L2 users was holding up the proceedings. The second was the instinct to express themselves with care in the L2 so as to avoid grammar errors and imprecise lexis. Participants were told that language errors were not a concern of the researcher, but they clearly found it difficult to set aside the prescriptions of their L2 instructors.

The brevity of what was written did not appear to be the consequence of a failure of auditory understanding. Indeed, during the retrospective verbal reporting that followed, participants tended to recall considerably more than they had covered in their writing, even without prompting by the researcher. They also tended to report coherently and logically, and the interview situation enabled the researcher to follow up the points made so as to establish whether the main propositions and the connections between them had been fully grasped. It became clear that writing imposed greater constraints than oral reporting, and that the summary task might even be seen as imposing heavier cognitive demands and additional skills such as the ability to précis.

On the evidence obtained, it seemed unlikely that the summaries would be informative enough to enable raters to form reliable judgements as to the lecture-listening skills of the writer. The conclusion was that verbal report was likely to prove a more valuable source than summary.

The research design was therefore revised. In the non-test condition, participants were still asked to take notes and to write them up, but these components of the task were used simply as prompts to assist the verbal report. Note-taking served an important role in reducing dependence upon memory and in simulating the real-life lecture situation, but it was also felt to be worthwhile to retain the summary-writing stage, since it enabled the participant to structure the information that had been obtained before presenting it orally to the researcher.

The proposal to assess lecture-listening skills by means of subjective ratings of written summaries was replaced by a more objective system of quantification based upon the number of macro- and micro-propositions accurately identified by the participant during the course of the verbal report. Further details are provided in Section 2.4.

2.3 Task conditions

Each participant was asked to undertake two tasks.

1 *Test-based*. They undertook an IELTS test from Section 4 of a past Listening paper, the section which aims to assess the candidate's ability to follow lecture-style material. Conditions were exactly as in the test: participants were given a brief period before listening to look through the questions, and were only allowed one hearing of the passage. The only difference was that the test was interrupted at certain points, when the researcher asked participants to report their answers and to attempt to give reasons for choosing them. All participants proved capable of reflecting and reporting on their own behaviour. The researcher followed up many of the explanations with requests for clarification or for further information; throughout, his attitude to the responses given was entirely neutral. At the end of the task, he asked respondents two general questions:

What was the main point or the main points of this talk?

Were there any parts of this talk that you found difficult to understand?
At the beginning?
In the middle?
At the end?

2 *Lecture-based*. In the second task, participants listened to a second Section 4 paper as if they were listening to a live lecture, and took notes with a view to writing a summary of what they had heard. They were allowed as much time as they wished to write the summary. They were then asked to report orally to the researcher on what they had understood of the interview. Like those in piloting, most summaries proved to be shorter than expected, and not as informative as the oral responses. However, this part of the task was retained because

   a. It assisted recall for the oral report.

   b. It gave participants the opportunity of representing the logical links between the various ideas in the talk and of assembling the information they had obtained before expressing it orally.

   c. It had some ecological validity in that it modelled what a university student might well be required to do when incorporating the content of a lecture into an assignment.

At the end of this task, the researcher asked the participant three questions:

What was the main point or the main points of this talk?

Were there any parts of this talk that you found difficult to understand?
At the beginning?
In the middle?
At the end?

Which of the two exercises did you find easier: the first or the second?
Can you explain why?

These last questions were followed up where necessary by a sub-question to establish more clearly if the perceived difficulty derived from the recording or from the task.

2.4 Materials and procedure

The two papers chosen for the study were taken from a recent collection of past papers (Cambridge ESOL 2005). They were Section 4 of Paper 1 in the collection (on the urban landscape) and Section 4 of Paper 4 (on the meshing of sharks in Australia). They were chosen because both had a similar relatively short running time and a similar density of informational content, and both featured a concrete but non-specialist topic. Question types were rather different; but it was felt to be important to control principally for listening content. The first recording is referred to as Text A and the second as Text B. The transcripts of the recordings appear as Appendix 1 at the end of this report and the task sheets for completion appear as Appendix 2.

The participants were divided into two groups. One group performed the first (test-based) task using Text A and the second (lecture-based task) using Text B. This is referred to as Condition A-B. With the other group, the order of texts was reversed; this is referred to as Condition B-A.

The two mini-lectures were transferred from CD to an iPod Nano for the purposes of the research. They were played to the participants through high-quality Bose Companion 2 speakers designed for iPod reproduction. The participant's verbal reports were recorded to computer using a Røde NT1-A studio microphone and digitised by a Roland Edirol USB UA25 interface. They were subsequently transferred to master CDs and then to cassettes to assist the transcriber.

Participants were explicitly told in the first task that they would be undertaking an IELTS test, but that the test would be paused from time to time for them to report, if they could, the reasons for choosing their answers. The pauses took place consistently after Questions 35 and 38 for Text A and after Question 34 and 38 for Text B. Before the second task, participants were told that they should imagine that they were listening to a lecture in a UK university and taking notes in order to write up a summary of the lecture.

All the ethical requirements of the University of Reading were met. The project was given approval by a departmental ethics committee, and each participant was asked to sign a statement of compliance before testing took place. Participants were paid £10 for their time.

The verbal reports were transcribed by a professional transcriber, using a format which numbers the lines of each report to ensure ease of reference. The transcription included not only the words of the participant but also any interventions by the researcher. To ensure confidentiality, participants were allocated letters in the order in which they were interviewed (from A to Z, then from AA to AC). As they appear in the transcription, each protocol has been coded according to the participant – the task – the text. For example, D2b refers to the protocol of Participant D when performing the second (lecture) task in relation to Text B.

Samples of two transcripts are included in Appendices 3 and 4, one in the A-B condition and one in the B-A. The two samples are from Participants R and V. They were chosen partly because these participants proved to be good at reporting on the processes they had employed and partly because the processes recorded were representative of those mentioned by the group as a whole. The participants counterbalance each other in that R was one of the Chinese sub-group, while V was European. Despite these differences in cultural background, there were certain similarities in their strategic behaviour. With V, the behaviour proved productive while with R it did not. R achieved one of the lower scores on Test A while V achieved the highest test score on Text B.

2.5 Population

The starting date of this research exercise was delayed until August 2006 to ensure the availability of suitable respondents. The project required naïve listeners – i.e. those without extended experience of residence or study in the UK. This was necessary in order to control for level of listening development: given that, once immersed in an L2 environment, different listeners develop at markedly different rates and in different ways. Furthermore, the experimental task entailed verbal report and thus required respondents to possess a level of English which enabled them to comment on aspects of their own listening behaviour. Of those available earlier in the year, a number were considered by their teachers to fall below such a level.

The population chosen for study was drawn from a group of students recently arrived to attend a pre-sessional course at the University of Reading. Intake in Reading is staggered, with the students possessing weakest proficiency scores arriving earliest. Participants were therefore chiefly drawn from the third (August) intake, on the grounds that they were not the highest fliers but that their speaking skills were likely to be equal to the task demanded of them.

All students in the August intake were circulated with a request for volunteers for the research study. The response was encouraging and sufficient to permit relatively careful controls to be applied in selecting participants. Volunteers were eliminated who had been previously resident in the UK for two months or more. In terms of first language, there was a heavy preponderance of students from the Far East. A decision was therefore taken to restrict to 12 the number of respondents whose L1 was declared to be Mandarin Chinese, of whom 8 were citizens of the PR of China and 4 were from Taiwan. In addition, to ensure a wider spread of first languages, a small number of students of European origin from the fourth intake were invited to participate.

Participants were chiefly limited to those whose listening scores on the University's own entry test ranged from 14 to 15 out of a maximum of 20 (IELTS 5.5 to 6 for those who had taken the exam). Speaking scores averaged 5.5 (IELTS also 5.5) out of a maximum of 10; they did not always correlate with listening scores. However, three participants were admitted whose scores showed them to be weaker listeners (10-12 on the Reading scale / 5 in IELTS) though their speaking scores suggested that they were adequate to the task.

The original proposal had been to base the study upon 20 students. However, student responses proved to be briefer than had been anticipated, with a typical session lasting around 50 minutes and the verbal report amounting to about 15 minutes per respondent. Data was therefore collected from 29 students in all. The wish to study greater numbers was prompted by emerging evidence of personal listening styles and processes. It also derived from the researcher's wish to ensure, so far as possible, that respondents were paired within first language groups, with one member of a pair performing in Condition A-B and one in Condition B-A. Clearly, L1 can be expected to affect the difficulty which a candidate encounters in a listening test – not simply in cases where L1 bears a phonological similarity to L2, but also in those (particularly European ones) where the two languages share a substantial number of cognates. Hence the wish to test speakers from a wide spread of first languages and the need to distribute first languages evenly across the two conditions.

Table 1 below lists the first languages of the respondents and shows how they were paired across the two conditions. By extending the numbers studied, it proved possible to group participants systematically, with only three anomalies (one Italian speaker grouped with one Portuguese, one Japanese speaker grouped with three Thai speakers; one Lithuanian speaker with extensive exposure to Russian grouped with three Russian speakers). The table also shows the balance that the study attempted to strike between respondents of Far Eastern origin and those of European origin. It had been the intention to feature speakers of Arabic and possibly Persian; but unfortunately those who were available had had previous periods of residence in the UK and had to excluded from the study.

Of the 29 participants, 19 were female and 8 were male.

Condition A-B		Condition B-A	
Student	**First language**	**Student**	**First language**
A	French	V	French
G	Portuguese (Braz.)	B	Italian
AA	Italian	AB	Italian
W	Greek	X	Greek
D	Thai	C	Thai
S	Thai	F	Japanese
I	Chinese (Taiwan)	K	Chinese (Taiwan)
L	Chinese (Taiwan)	M	Chinese (Taiwan)
N	Chinese	O	Chinese
P	Chinese	Q	Chinese
R	Chinese	U	Chinese
T	Chinese	Z	Chinese
AC	Russian / Turkmen	J	Russian / Turkmen
Y	Russian		
E	Lithuanian		
		H	Nepali

Table 1: Paired participants showing first language

The results for the one unmatched speaker (of Nepali) are omitted when comparing performance across the two conditions (and particularly when comparing them numerically). The results are also omitted for Participants E and Y (the Lithuanian – Russian pair) as they both fell into Condition A-B. This leaves 13 pairs, which include a block of 6 pairs of native speakers of Mandarin. Within this block, it was felt to be important to distinguish between those originating in Taiwan and those originating in Mainland China because of their different educational backgrounds and traditions.

3 DATA ANALYSIS

The data analysis adopts three main lines of enquiry.

a It compares the marks achieved in the formal IELTS test with a quantification of the extent to which participants were successful in extracting information from a mini-lecture in non-test conditions.

b It examines evidence from verbal report of the means by which answers were achieved in the simulated IELTS test; and distinguishes between processes specific to the test condition and those which might also occur in a less constrained experience of lecture listening.

c It examines verbal reports by participants comparing the test and the non-test conditions, to establish how different they perceive the underlying processes to be.

3.1 IELTS score

The tests were first marked by reference to the answers specified by the setters. To ensure maximum reliability, the exam board's regulations require strict adherence to these forms, in terms of both wording and spelling. However, this stipulation would have disqualified a disturbingly large number of the participants' responses (19 in total) which strongly indicated that full understanding had been achieved. The items in question were as follows:

Text A

Q 35 less *dangerous*.

1 instance of *danger*.

Q38 *considerably reduce / decrease / filter* (the wind force).

5 instances of *break* 1 instance of *reduce*

Text B

Q34 (Sharks locate food by using their) *sense of smell*

10 instances of *smell* 2 instances of *nose*

A limited range of answers is provided to markers in these tests; but it is curious that variants like those evidenced here did not occur during piloting by Cambridge ESOL. *smell* is clearly not as elegant as *sense of smell*, but is surely acceptable given that candidates' written expression is not at issue. As for *nose*, it fits the context perfectly adequately. The entire chunk *break the wind force* actually appears in the recording in a slightly different form (*break the force of the wind*); given this, it seems unfair to rule it out as a possibility. A check of the verbal protocols showed that in all of the cases cited the respondent had achieved a full understanding both of the tenor of the question and of the relevant information from the recording. On these grounds, the variant responses were accepted for the purposes of this study.

With this adjustment, the results recorded for the two tests ranged from 6 to 10 for Text A (N = 15) and from 5 to 9 for Text B (N = 14). The respective means were 8.2 (SD 1.37) and 7.07 (SD 1.07). The spread of marks was unexpectedly wide, given that all respondents except three had achieved very similar IELTS and / or Reading Listening scores.

3.2 Lecture-listening competence

In the lecture-based task, the protocols obtained from participants consisted of free recall of as much as possible of the mini-lecture that had been heard, prompted by the notes and summary that had been written. An objective means was sought of establishing what proportion of the available information was reported by each participant. To this end, it was necessary to identify the different points that were made by the speaker – but to do so in a way that was sensitive to the relative value of those points and to their contribution to the overall discourse structure.

The two target texts were rather different in structure. Whereas the first featured one overriding topic (the role of trees in urban planning), the second embraced two (the characteristics of the shark and the use of netting to protect bathers). Within those topics, a series of macro-propositions were identified in Text A, based upon the paragraphing which the setter had used when transcribing the lectures. Within the paragraphs, a set of micro-propositions was then identified. Text B was treated simply as a series of micro-propositions

The topics and propositions for each recording were listed with no indication as to perceived importance. They were submitted to five judges with extensive experience of ELT (and particularly of the teaching of discourse for EAP), who were asked to grade them in relation to the texts as 'macro-', 'micro-' or 'peripheral' (i.e. at a low level of importance to the text as a whole). Their feedback was then compared and collated to form profiles of the content and discourse structure of the two recordings used in the study. The profiles appear in the two panels below.

RECORDING 1: Trees and the urban environment

MACRO-PROPOSITIONS

1 Trees change climate

2 Trees regulate own temperature

3 Trees reduce the strength of winds

4 Trees reduce traffic noise

5 Problem: trees need space

MICRO-PROPOSITIONS

1a less windy
1b cooler
1c more humid
1d less dangerous

2a water through leaves
2b trees cooler than buildings [buildings 20% more than human temp]
2c trees humidify the air

3a high buildings produce winds at ground level
3b trees filter the wind

4a BUT much vehicle noise goes through trees
4b BUT low frequency noise goes through trees

5a roots and branches
5b difficult to plan in a narrow street
5c water, sunlight, space

The procedure thus in many ways adhered the 'macro' / 'micro' principles of Van Dijk and Kintsch (1983), but with added validation obtained from:

a external judgements as to the relative importance of the propositional information

b the neutral decisions made by the original transcriber when dividing up the content of the mini-lectures into paragraphs.

At one point, an attempt was made to represent the complex hierarchical relationships between propositions along the lines of Gernsbacher's (1990) Structure Building model. However, the exercise proved too complicated for practical purposes. It was also recognised that in the informal conditions of verbal report (the main source of data), participants could not be expected to mark inter-propositional relationships as unambiguously as they might in a written summary.

RECORDING 2: Shark meshing

TOPIC 1 Characteristics of the shark

MICRO-PROPOSITIONS

1a Large: length [10-16 metres]
1b Large: weight [795 kg]
1c Flexible skeleton
1d Barbs not scales
1e Quick swimmers
1f Fins and tail
1g Keep swimming unlike other fish
1h Bottom of ocean
1i Food on ocean floor
1j Sense of smell

TOPIC 2 Shark meshing

MICRO-PROPOSITIONS

1a Large nets parallel to shore
1b Set one day, taken out to sea the next

2a Began 1939, only Sydney
2b 1949 extended [beaches to south]
2c 1970 Queensland

3a NZ and Tahiti – no
3b South Africa – yes

4a 1500 first years
4b 150 per year now
4c caught in warmest months [active when air/ocean at max temp] [Nov-Feb]

5a NOT sharks unafraid
5b NOT sharks biting holes
5c waves and currents
5d sand moving, can't hold nets

The protocols for the lecture-based task were then analysed to establish how many propositions each participant had reported. The number of relevant propositions identified was taken to constitute evidence of how much information an individual listener had succeeded in extracting from the text. For Text A, the count included those identified as both macro-propositions and as micro-propositions. For Text B, micro-propositions only were counted. Also calculated were the number of propositions incorrectly reported and the number of peripheral items of information included (an indication that the main argument had not been followed).

The results were tabulated alongside the scores obtained by participants in the earlier administration of the IELTS test. They appear in Tables 2 and 3 below. Participants are ranked by their scores in the test-based task, shown in the first column of figures.

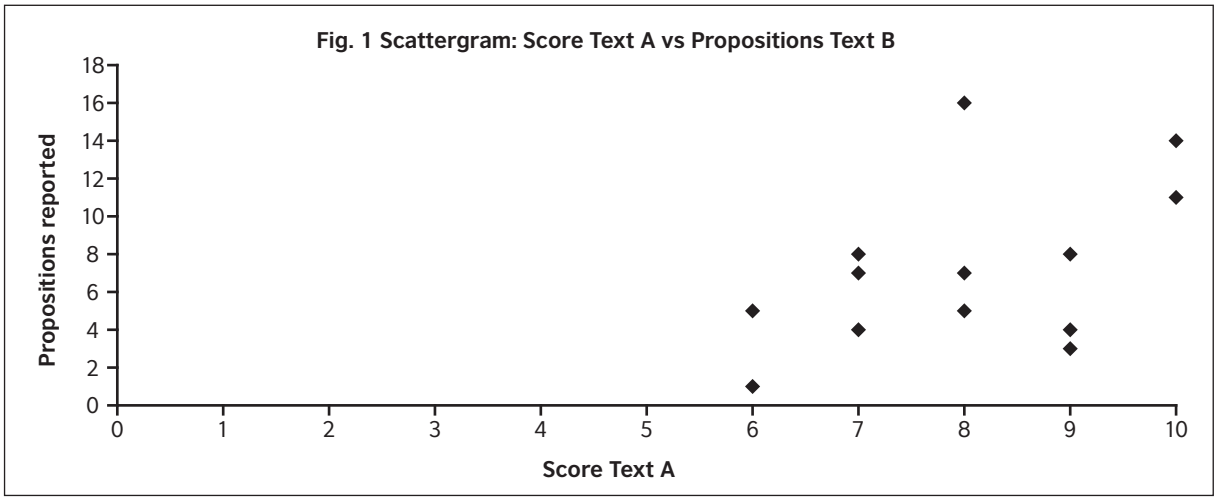
Participant (N = 13)	Test score (A)	Correct propns (B) (Tot = 24)	Incorrect propns (B)	Peripheral info (B)
A	6	1		
P	6	5	4	2
D	7	8		
R	7	4	4	1
T	7	7	2	
S	8	7	1	
W	8	5	1	
AA	8	16		
G	9	8		3
I	9	3	1	3
N	9	4	5	
L	10	11	3	
AC	10	14	1	
Mean	8.0 (1.35)	7.15 (4.34)		

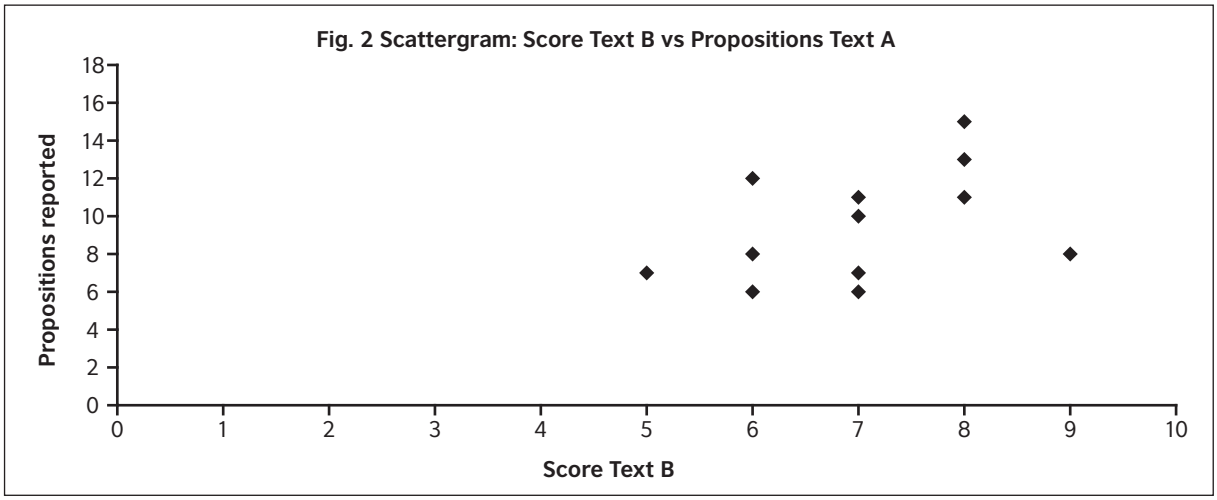
Table 2: Test scores (Text A) versus evidence of successful lecture-listening (Text B)

Participant (N = 13)	Test score (B)	Correct propns (A) (tot: 19)	Incorrect propns (A)	Peripheral info (A)
Z	5	7	1	
C	6	6		
F	6	8		
AB	6	12		
B	7	11		
J	7	7	1	
K	7	10		
M	7	6		
O	8	13		
Q	8	15	1	
U	8	13		
X	8	11		
V	9	8		
Mean	7.08 (1.11)	9.77 (2.98)		

Table 3: Test scores (Text B) versus evidence of successful lecture-listening (Text A)

As part of the cognitive validation exercise, evidence was sought of a statistical correlation between the figures for correct answers in the test and the figures for number of propositions reported. The two conditions (A-B and B-A) were treated separately. For A-B (test based on Text A and propositions reported from Text B), the Spearman rho correlation was rs = 0.43, N = 13, n.s. For the B-A condition (test based on Text B and propositions reported from Text A) the correlation was rs = 0.53, N=13, n.s. In neither case did the statistic indicate a significant relationship between the score obtained on the lecture-based section in the IELTS paper and the ability to report propositional information from a recording of a lecture heard under non-test conditions. The lack of correlation between the two measures is confirmed by the scattergrams in Figures 1 and 2.





In considering this evidence, the cautionary note sounded earlier needs to be repeated: the sample was a relatively small one (inevitably, given the method employed). A normal distribution cannot be ensured, and these results should be treated as broadly indicative rather than conclusive.

In these circumstances, it is worthwhile examining individual cases. The two participants who scored lowest in the IELTS test based on Text A also showed signs of difficulty in unconstrained lecture listening. Participant A only succeeded in reporting one proposition, and, while P reported 5, she also misreported 4 and included information that was not central to the topic. Similarly, Participants L and AC who achieved scores of 10 in the test were also among the three highest in terms of the number of micro-propositions reported. These cases do indeed suggest some relationship between the aspects of the listening construct tested in the IELTS section and the skills demanded by a more ecological lecture-listening experience.

Of course, the possibility remains that what is tested in the IELTS paper is a general listening construct which has relevance to all listening tasks, rather than any trait specific to lecture-listening. Furthermore, the examples cited are not entirely supported by the results from the B-A condition, where the participant with the highest test

score only succeeded in identifying the same number of propositions as a participant with the second lowest. The participant who was most successful in the lecture task only achieved a score of 8 in the test. Perhaps the most interesting anomaly lies in Participant AA from the first group. She achieved a score close to the mean on the test and had a low IELTS Listening score of 5 from earlier testing; but she produced an impeccably detailed and accurate account of the lecture.

Results for the subset of eight participants from Mainland China were examined separately and compared with those for the participants as a whole. The mean scores for both Test A and Test B were 7.25, close to the overall means. Curiously, members of this group showed themselves markedly less able than others to extract propositional information from Text B (mean number of propositions = 5, as compared to 7.15) but markedly more able to do so from Text A (mean number of propositions = 12.5, as compared to 9.77). Given the small numbers, too much should not be made of this finding.

Clearly, limitations of linguistic knowledge and listening competence restricted the information that the participants were able to derive from the text. A further possible explanation can be found in the additional cognitive demands associated with processing input in a second language. The need to focus greater attention upon word recognition and syntactic parsing potentially imposes restrictions upon the amount of information that can be processed, stored and/or recalled. It is entirely understandable therefore that relatively few micro-propositions were recorded by some listeners. That said, their ability to identify larger topics in a text like Text A (at least at the stage where the mini-lecture was summarised) seemed to be relatively unaffected by processing difficulties or demands at local level.

3.3 Test-specific behaviour

A second line of enquiry examined the protocols relating to the test-based task, for evidence of how participants had arrived at the answers they had given. Each participant had been asked not only to report each of their ten answers orally but also to provide a rationale for having chosen it. The purpose of studying these rationales in detail was to identify to what extent the processes employed by test-takers conformed to those that might be applied in real-life lecture-listening situations and to what extent they took advantage of the additional information available in a test and/or explored strategic routes that were specific to the testing context.

3.3.1 Cognitive validity of the tasks

It is worthwhile at the outset to take note of the differences between the information available to a candidate taking the two IELTS listening sections that formed the basis of the study and the information that might be available to a participant in a typical academic context. It is also worthwhile to draw some general comparisons between the requirements imposed by the test methods and those that language processing research tells us obtain in real life listening contexts.

The task sheets for completion by candidates form Appendix 2 of this report. The task for Text A consisted of a note completion exercise that is much favoured by IELTS setters in the Listening paper - presumably on the grounds that it achieves face validity by resembling the type of note-taking that might take place in an authentic context. The first part of the task for Text B consisted of a similar note-completion exercise. It was followed by four multiple-choice items of four options and a further one in which two options had to be chosen out of seven.

The note-completion task for Text A provides a strategically minded candidate with the following *gratis* information before even hearing the recording:

- an outline of what the lecture covers, with some lexical gaps
- a set of gaps to be filled that closely follow the sequence adopted by the lecturer (some even forming part of a list).
- key words by means of which to locate information in the mini-lecture
- one constituent of two relatively frequent collocations: *ground level, low frequency*
- two sequences which reproduce the oral text word for word, with one word omitted.

Interest here attached to the test-wiseness of the participants in the study and the extent to which the protocols showed that their answers were influenced by this externally provided information rather than by the evidence of their ears.

The gapped notes for Test A are quite detailed – raising issues of whether validity is compromised by a task that taps in to the reading skill to such a degree. In process terms, the level of detail and the organisational structure of the notes mean that the candidate is not required to undertake certain critical meaning building operations

which would normally play a central part in lecture listening. These include (Brown 1995; Field 2004a, pp 163-5; Field 2008a, pp 241-265):

distinguishing main points from subsidiary ones (though admittedly this function may be provided by a handout in a real-life lecture context)

distinguishing new propositions from instances of rephrasing and exemplification

recognising the argument relationships that link propositions

integrating incoming information into an ongoing discourse representation.

The focus of the testing, in other words, is very much 'bottom-up' in that what the candidate has to contribute chiefly takes the form of lexical matching. In this respect, it is difficult to see that it replicates the range of EAP processes for which the test aims to serve as a predictor.

The similar task for Test B was in a much more abbreviated form, imposing a lighter reading load. The answers were to some extent predictable using topic knowledge; though similar strategic behaviour might well be employed by a listener in a non-test context. The multiple-choice options were mainly brief and some required the candidate to assess the status of two or more pieces of propositional information rather than simply performing lexical matches (for example, recognising the negative attached to *Tahiti* and *New Zealand*). The exception lay in questions 39 and 40, where key words (*strong waves and currents, moving sands*) closely echo the recording. On this analysis, one might say that the tasks set for Text B appear to achieve greater cognitive validity than those for Text A, and that one might expect less evidence of test-specific strategies.

3.3.2 Evidence from protocols

Two lines of enquiry were adopted. A distinction was made between responses which indicated that the participant had relied upon the written words in the task items in arriving at an answer (in other words, a listening process driven by reading) and those which suggested a primary reliance upon the spoken signal.

Use of written information

There was extensive evidence of participants adopting a procedure of matching information from the written task sheet against what was heard in the recording. The cues that were used seemed very often to be at word level rather than at propositional level. The listening process was partly shaped by a strategy of scanning the recording for words which resembled those in the items or were paraphrases. This attention to word level was sometimes at the expense of wider meaning. One participant, who scored the mean of 8 in the test, was candid about the way he focused his attentional resources:

(1) *[the main point was] preserve tree but I'm not quite sure because + every every time I use + I mean my my method to + listen to to do the IELTS listening + yeah I just look at the words not focus what it is about* (S1: 145)

[Here and throughout, quotations from protocols are referenced by the participant's code (here S) followed by the figure 1 or 2 (indicating first or second task) plus a reference to the line in the protocol where the extract begins. To separate citation from main text, the participant's turns are italicised, while the researcher's interventions are shown in a non-italic font. The reverse is true in the database].

The scanning strategy was supported by the convention that items follow the same sequence as the text and a widely-shared expectation that items would not occur too closely together. The latter feature is entirely reasonable in that the candidate needs time to record an answer; but participants showed themselves aware of the strategic possibilities afforded by the feature of the test:

(2) *so when I was reading the answering the first one + she was maybe she had already finished the list no? + the other case is even if the words maybe were I made some mistakes in other parts I mean + but you have time to write to listen because when you were + when I was writing er + she was speaking about something else not important for the test.* (AA1: 148)

There was evidence from the protocols that some participants used the spaces between pieces of targeted information to switch their attention back to the written task sheet in search of possible cues to the next item to come. This became apparent when several of them admitted missing information in the recording because of excessive attention to the written material.

(3) *I missed it because I didn't I didn't + I didn't realise the 'frequency' has came so quickly* (P1: 99)

(4) *er when I try to get this answer um + he he is already talking about the make cities cooler yes + so I missed the answer* (T1: 19)

The text-to-recording strategies varied from participant to participant and from question to question, but were classified as falling into four main types. As already noted, they seemed to operate principally at lexical level, with single words, lexical phrases or potential collocations used as cues.

a The respondent used a word or words from the written text as a means of locating information in the spoken text. [Q loc]

(5) *and maybe also for the wind force when I hear two er + two thing er+ two different level + ground and high+ and so it's um + I don't know which one is good because er + with just looking about something before 'level' + and if we have two + twice 'level' it's + it's confusing a bit* (A1: 39)

b The respondent listened for words in the spoken text that formed a one-to-one match with those in the written. [Q match]

(6) *yes because she introduced the um + er wind effect on buildings so er + when I heard this word 'buildings' + 'wind effect' 'wind force on buildings' + so I concentrate um + she perhaps she followed the the question written* (AA1: 85)

c The respondent sought a paraphrase in the spoken text of a proposition expressed in the written one. [Q para]

(7) *yes I because 'coastline', 'beach' er + is + are very similar so um + I don't know + the meaning is quite the same . . .* (B1: 37)

d The respondent chose an answer according to its position in a list or in a sequence of propositions in the written test. [Q seq]

(8) *and her er + some the recording give some some interrupt er + because er + he she said she 'water' before 'the sunlight' + but at end is the room* (R1: 107)

Table 4 shows the strategies reported by participants in respect of the two tests. The most common strategies were widely generalised across participants, with only two out of 13 failing to record a Q loc for Text A and two out of 13 for Text B. No participant recorded more than 4 occurrences of the same strategy across the 10 items – suggesting that their use reflects the form and demands of the item.

	Q match	Q loc	Q para	Q seq
Test A (N = 13)	26 (18.98%)	2 (1.46%)	15 (10.95%)	
Test B (N = 13)	30 (22.56%)	25 (18.80%)	2 (1.50%)	3 (2.56%)

Table 4. Test-wise strategies employing visual cues.
Percentages based upon a total of 137 instances for Test A and 132 for Text B. The totals exceed the number of items answered (130 in each) as two processes were sometimes cited as having contributed to a single answer.

The finding that candidates make use of test item wording in achieving their answers will cause no surprise – though it raises concerns for construct validity since the cue employed engages a different modality from the construct that the test aims to measure. But it is striking here how many of the participants' answers were achieved by these means. The table shows that strategies generated by written input in the form of test items were cited as instrumental in achieving around 40% of the answers given in Test B and 30% in Test A. It is also apparent (despite the initial impression recorded above that Test B was more cognitively valid) that the multiple-choice format employed in B promoted a greater level of test-wise strategy use. Several of the participants attested to the way in which the MCQ options had led them to seek the spoken forms, associates or synonyms of key words which they had seen in written form. In cognitive terms, the MCQ format could be said to promote a process of checking information against pre-established cues rather than the more ecological one of receiving, interpreting and organising it.

The issue so far as cognitive validity is concerned is that these channels for establishing meaning would not be available to the participant in a natural lecture setting. Admittedly, their role might to some extent be paralleled in the provision of a handout; but the process of matching handout propositions to those in the presentation is a somewhat different one. Handout material is unequivocal (as against the multiple options offered by MCQ), is fully formed (as against gapped notes) and constitutes, in effect, an abbreviated paraphrase of the spoken text. The process in which the auditor engages entails tracking from proposition to proposition rather than from key word to proposition.

Auditory word matching

A second group of processes seemed to be more reliant upon the auditory signal. In distinguishing these instances from those in Table 4 (identification on the basis of a written cue), it was not easy to determine what role, if any, the wording of the test items might have played . One must assume that, at the very least, the identification process was assisted by an awareness of the word class and lexical field to which the target item belonged.

Here again, the focus of attention seemed to be very strongly at lexical level. The rationale supplied for an answer was very frequently 'I heard the word'. Figures indicating the extent to which lexically based techniques were employed appear in Table 5.

	Lexical recognition	Cognate in L1	Collocation recognised	Phonological transcription
Text A	60 (43.80%)	3 (2.19%)	5 (3.65%)	10 (7.30%)
Text B	31 (23.31%)	0	2 (1.50%)	0

Table 5. Reported lexically-based strategies employed by participants

Four processes were identified: matching to a known word, matching to a cognate in L1, identifying a two-word collocation and attempting the transcription of an unknown word using phoneme-grapheme correspondence rules. It sometimes proved difficult to establish if the lexical matches entailed a full grasp of the surrounding context; but in some cases it was made clear by the participant that it did not. Here are some examples where understanding above the lexical level was not achieved, but was sufficient to achieve a correct answer:

(9) What does low frequency mean?

++ it means er + the trees can reduce the noise er + because I think the noise goes through the trees . . .

but what does 'low frequency' mean?

low frequency mean ++ not er frequently (T1a: 141)

(10) *. . . answer number 36 is 'leaves' and er + I don't know its exact spelling + un but OK was clear because er + it's the last word of the sentence...*

OK + what did they say about leaves?

er I don't remember + I was concerning to wait for the the next point ... I was actually paying paying attention to the building (AA1: 64)

(11) *humid yeah + probably um + what's the nature of tree like er + um + how how are how do trees er grow and um + something like that* (S1: 169)

The much greater incidence of lexical targeting in relation to Text A would seem to indicate that test method was an important factor, with the gap-filling exercise encouraging candidates to direct attention at word level. As for Text B, 13 of the 31 instances recorded (9.8%) were cases where either the gap or the MCQ option demanded a number, and participants admitted to simply scanning the upcoming section of the text for numbers.

Here, conclusions on cognitive validity need to be hedged. On the one hand, lexically based lines of attack appear to be very common among L2 listeners whose understanding of a listening passage is less than complete (Field 2004b, 2008b). They would certainly be employed in the circumstances of listening to an actual lecture. On the other hand, the 'bottom-up' nature of the gap-filling testing method must surely play some part in directing additional attention to local, word-level processing. Because human attention is a limited resource, a processor needs to be selective in the information it retrieves (Styles 2006). This may explain why some of the participants reported having located a correct word without understanding its significance to the lecture as a whole. The problem was not one of general comprehension, but one of having directed attention in 'spotlight' fashion to the word or phrase which supplied the necessary answer, leaving insufficient capacity for wider considerations.

Other processes

Table 6 provides a summary of the remaining processes reported by participants. With the exception of the use of world knowledge, they featured in the reported behaviour of only one to three participants.

	Miss	Random	World	Prom.	Discourse	Elim.
Text A	10 (7.30%)	1 (0.73%)	4 (2.92%)	1 (0.73%)		
Text B	5 (3.76%)	6 (4.51%)	18 (13.53%)	0	2 (1.50%)	9 (6.75%)

Table 6. Additional processes reported by participants

Some of these processes do indeed play a part in successful useful lecture-listening skills. For example, many participants used world or topic knowledge with Text B. It quite often led them to wrong conclusions in the test condition - though interestingly it did so considerably more often in the non-test condition where they had to construct a meaning representation from scratch:

[they] try to suggest how do we + er how do they er + preserve the shark [S2: 32-33]

how about they attack humans + on the beach, swimming beach [P2: 27-28]

and the shark only live in the warm water um [R2: 22]

he said that sharks were not so dangerous [Y2: 29]

Much is made in the literature of the value of topic knowledge in supporting comprehension (Long 1990; Schmidt-Rinehart 1994); but this data indicates that its effect can also be counter-productive, and lead to second-guessing.

Curiously, two techniques for lecture listening which are much discussed in EAP listening materials (such as Lynch 2004) – namely, paying heed to prominent items ('Prom' in Table 6) and to discourse markers ('Discourse') – were little reported.

Other processes appear to be test-specific. 'Miss' records instances where participants missed the information because their attention was directed elsewhere. The cause was often explicitly related to the demands of the test: the need to spell correctly and check previous answers or (as already noted) a strategy of closely following the order in which test items occurred. 'Elim' indicates cases where the participant reached a conclusion as to the most likely MCQ option by rational consideration of the evidence. The types of cue used were: most likely option, the option on which most was said and the last-mentioned option (on the reasoning that speakers often reject several points before accepting one). Again, here the participant's behaviour seems to be chiefly driven by the format and demands of the test.

The processes used by the L1 Mandarin group to achieve answers were examined separately. There was extensive use of lexical identification and of the strategy of using item wording to locate information in the text. While these patterns of behaviour were not dissimilar to those of some other participants, what was striking was how consistent they were across all members of the subset. These participants appeared to be more consistently strategic as a group – or possibly had been trained to be so.

3.3.3 Ability to identify 'main points'

The apparent bias towards processing at word level that had been evidenced in the protocols for the test-based condition was investigated by studying participants' responses to the question: *What do you think the text was about?* [Four participants were not asked the question directly, but their responses indicated quite unambiguously whether they had or had not grasped the main points].

The responses were analysed using very narrow criteria. Participants were credited with having understood the 'main points' of the two tests if they specifically mentioned:

For Text A: *cities / urban environment trees*

For Text B: *description of sharks protecting beaches from sharks*

A score was given for each point mentioned.

One hypothesis was that listening to the text under test conditions might have diverted attention from the main points. An alternative hypothesis was that listening under non-test conditions without the support of the task sheet might have made it much more difficult for learners to identify the main points.

Results were totalled for 26 participants, of whom 13 had heard Text A under test conditions and 13 had heard Text B. Means number of scores were calculated (out of a maximum of 2 per text) and appear in Table 7 below.

	Text A Test	Text A Non-test	Text B Test	Text B Non-test
Mean (max 2)	1.54	1.77	1.15	1.38

Table 7. Main points identified: means across non-native participants

With the usual rider about the smallness of the sample, it would seem that this population was more likely to identify the main points of the lecture in a natural situation than in a test one. The test method may have served to distract attention from the main points - despite the fact that test-takers held evidence of the speaker's intentions in the form of a worksheet and that many of them reported paying close attention to this information. One conclusion is that that the participants' attention was quite closely directed to the wording of the worksheet as part of a test-wise strategy of key word matching, at the expense of wider meaning.

4 REPORTED EXPERIENCE OF PARTICIPANTS

4.1 Relative cognitive demands

A third and final line of enquiry examined participants' perceptions of the relative difficulty of the two tasks, and the reasons they gave for their views. As reported above, all participants were asked *Which of the two exercises did you find easier: the first or the second?* They were also asked to explain their choice.

A working hypothesis was that most of them would respond that the non-test condition (note-taking and oral report) had proved more demanding than the test condition. The reasons for adopting this assumption were as follows

a As discussed in Section 3.3.1, a test candidate benefits from supportive information in the wording of the test items. The information was not available in the note-taking condition.

b The non-test condition was more complex in that it required the participant to engage in three processes: note-taking, summary writing and oral report. It also demanded a longer attention span in that the recording was only paused once and very briefly. On these grounds, it was assumed that the task was considerably more cognitively demanding than the test items and would be reported as such by the participants.

If this hypothesis were true, one might expect participants to respond consistently that the first task was easier than the second. In other words, in the A-B condition, they would report their experience of processing Text A to have been easier than that of processing Text B (A < B). They would report the converse in the B-A condition (B < A). This finding would raise issues of cognitive validity in that it would suggest that the processes elicited by the test were less cognitively demanding than the real-life ones for which the test is intended to serve as a predictor.

Participants' responses to the 'Which was easier?' question are shown in Table 8 below. No response is recorded for Participant A as she was not asked the question. It is evident that the hypothesis was not supported in any consistent way. Especially striking is the extent to which the responses varied according to which test had been taken. The majority of respondents indicated that their experience of taking Test A was less demanding than the subsequent note-taking task. Even here, two participants (S and AA) dissented, while one respondent (E) indicated that she found the two tasks equally demanding. By contrast, when the material presented under test conditions was B, the majority of the responses (by 9 to 5) indicated that the respondents *had found the note-taking task easier than undergoing the test*.

Two important differences distinguish the two IELTS tests that were chosen for this study. Firstly, the recorded material in A is less detailed and less propositionally complex than the material in B (see panel in Section 3.2). Secondly, while the test method in A consists entirely of gap filling, in B it is a combination of gap-filling and multiple-choice.

Text A in test condition		Text B in test condition	
D	A < B	B	A < B
E	A = B	C	A < B
G	A < B	F	B < A
I	A < B	H	A < B
L	A < B	J	B < A
A	A < B	K	B < A
P	A < B	M	B < A
R	A < B	O	A < B
S	B < A	Q	A < B
T	A < B	U	A < B
W	A < B	V	A < B
Y	A < B	X	B < A
AA	B < A	Z	A < B
AC	A < B	AB	A < B

Table 8. Participants' reports of relative task difficulty

The researcher was at pains to ask participants to consider the difficulty of the task separately from that of the recording (indeed, two participants, Q and X, actually gave different answers for task and recording). Even so, one has to recognise that some participants may have found it difficult to separate the two when reporting on the cognitive demands of the exercise. But an alternative conclusion is that at least some participants who reported A < B were influenced by the heavy attentional demands made by the multiple-choice format. In other words, it was not so much that note-taking was easy as that the demands imposed by the test were hard.

4.2 Protocol evidence

4.2.1 Views on note-taking

Further insights were obtained by examining in some detail the comments of participants on the two tasks. Here, an entirely unexpected finding was that 8 of the 28 participants questioned (28.5%) categorically asserted that they found the process of note-taking easier than operating under test conditions. Some extracts follow to illustrate the points that were made in support of this argument.

(12) did you find it more difficult to take notes or to answer the questions?

I think sharks + um sharks is more difficult to answer the question + because the question is is + how do you say that? um + um in the sharks there are some questions I can't catch it exactly.

so the questions make you listen for things?

yes.

and sometimes you don't hear them?

yes + so I can guess it.

so when you're taking notes you don't have to listen for anything.

yeah.

you can just write down what the lecturer was saying?

yes and I can er + from the stress I can know which is much important.

right from + from what the lecturer says?

yeah.

so it's easier to do it when you don't know what you are going to hear...

yeah yeah. (Q2: 40)

(13) OK + what about writing the notes and everything? + was that difficult?

um yeah + difficult to write to find what is main point about.

right.

but I think it's not difficult if we we try to get information + but + er + I don't know if it necessary or not.

right + OK + but the notes that you wrote...

yeah.

. . . were they more difficult than filling in the um + answers for the urban landscape?

yeah, they're easier.

you think it was easier to write notes?

easier to write notes. (S2: 55)

---

(14) well the last question is + did you find this more difficult than the last one with the sharks or not?

I think er why do the question is more nervous + and

the questions make you nervous, do they?

yes make me nervous and er + maybe I have readed the questions + sometimes maybe questions can give some information about the + what they will say + but the questions it's they er + more interesting in some numbers or some words er+ maybe I can I am not understand the words and

and that worries you?

yes that worries me.

so did you find the shark one more difficult to do than writing notes?

yes because I have + I have to read some questions and + that does use a lot of my account attention + and I cannot put so many attention on the context it said + such as the questions + OK this questions ask for something + and I just wait for the question and do not listen to others. (U2: 54)

---

(15) Which did you find more difficult?

the first one + definitely.

why?

because um with this + this tests it it is um + how can I? + it's + I have I have a lot of more stress with this sort of test because you um + you don't want to miss any answer + but with this technique it's + it is different + um even if you miss something er + you you you will understand the the general idea what is talking about + but in this test sort of test um + you when you miss miss the point + then you get you get stress and then for er+ for the following questions it's harder + and so it's quite difficult.

do you think you were behaving differently as a listener in the second one? Were you doing things that were different from the first one to the second?

if if what? + I'm sorry.

were you behaving differently as a listener when you did the first test + to the way you behaved in the second + were you listening in a different way?

um yes definitely +there's not the same way + yeah +because in the first part I'm just focusing on words not the general meaning + but in the second I focus on the the the general ideas + the most interesting points. (V2: 26)

---

(16) why did you find [the second task] easier?

um maybe I can focus on the um + the the the the lecture + um she said um + what really is important + and can summarise + but I if I heard the the + and deal with some question sometimes I feel nervous + and just focus out or + catch the catch the word.

so the questions make you nervous?

yes.

but you don't feel nervous when you're listening

yes lecture.

to a lecture.

yes + yes.

OK + didn't you find it difficult to take notes + and to understand what the person was saying?

no + I think er taking notes it's better.

taking notes is easier for you.

yeah is easier for me.

do you think you understood more?

yes + understood more. (Z2: 62)

---

(17) um you've done an exercise on the urban landscape and one on the sharks + which did you find easier?

um + maybe this one.

the one on the sharks?

yeah.

why?

because er + when I take my notes I can + I don't pay attention to my er spelling + I use abbreviations symbols something + that after if I have the time I can recognise a symbol or something + in that case I have + I think that I will + I would have been marked also for the spelling.

oh you mean with the + with the urban landscape?

that one + yes + so er and um + I don't know + I think it's easier because you you + there you have something ready filled + already filled out.

yeah, I see + this is the urban landscape?

something you know you have also to check out + before what's filled out what's not + and er + here er was my job + I mean I know what I'm going to write + I recognised the the key key words and whatever +, what else er + he said + I these key words make me remember all the rest.

yeah.

or +

which one do you think you understood better?

this.

the shark one.

yeah I didn't really + I didn't care + when I if I have to fill out only some particular sentence+ a word in a sentence + I pick actually that sentence + and I wait to listen on some words that are OK. (AA2: 51)

---

(18) which did you find harder to do + answering the questions or taking notes?

er ++ hard to do + to take notes or answer the question?

yeah.

(laughs) It is + it is different strategies because er + I I'm + my + generally I take notes so for me perhaps it's more simple + but other people perhaps er + it's better to read the the question + the . . .

what did you find?

me fine to take note because I usually take note + but perhaps if you er + know that the IELTS test is based on the question + you can er + learn to read quickly the question + and then these are different strategies I think but +

so you found it difficult because you had to read the questions and so on yeah?

sometime to read the question is better because you can er + predict of you have to listen + but er I I

don't either pref + for example if I have to do the IELTS test + I I can improve my excuse to read the question then to answer + in my example er for me it's more simple to take note because er + I'm just usually take notes during a lecture not to read the question + but er if you have only to take a test er + perhaps it's better to have the question then + as well to answer the question + OK. (AB2: 46)

These protocols have been quoted at length because they serve to highlight two important findings:

a Some participants (Q, U, V) appear to feel that, under test conditions, the process of storing questions in the mind and scanning the recording for appropriate answers imposes considerable task demands. Indeed some (U, V, Z) claim that it causes stress. AA mentions the additional demand (irrelevant to the listening construct) of ensuring accurate spelling.

b Some participants (V, Z, AA, AB) report that they listen very locally and at a level of minor detail under test conditions but much more globally when note-taking.

On these grounds and contrary to expectations, the participants quoted found note-taking *less* cognitively demanding than operating under test conditions.

4.2.2 Support provided by the task sheet

Quite a large number of respondents took the opposite view: that working under test conditions is easier. As predicted, a number of them represented note-taking as imposing greater cognitive demands. Participant M put it as follows:

(19) *er this was more difficult I think because it's er + ser yeah a lot of serious speech in this detail + and she didn't stop in each part+ and we have to er + summarise in our mind every part of his speech and to organise again + because some some of her idea is the + jump + this is for + this part is problem*.

so you think she + her ideas moved very quickly?

quickly.

from one point to another one?

and she the point + yes where we have to summarise in my mind and try to write down the summaries + they more difficult + and we + direct questions I don't have to + the the end the question yeah + so I can follow the question to find some details. (M2: 47)

What M is drawing attention to in her first turn is a major difference between the test taker and the note-taker: the latter has to establish the relative importance of the points that are made by the speaker and to construct an overall discourse representation (see Gernsbacher 1990 on the complex demands of building argument structures). In the circumstances of a test, either the setters tend to target a series of points regardless of their relative importance, or the task sheet provides the candidate in advance with an argument framework for interpreting the recording. A similar point is made by S in the quotations in the previous section, though he still maintains that note-taking is easier.

The support provided by the test paper is mentioned by many of those who felt that note-taking was more demanding. A point frequently made concerned the fact, explored in Section 3, that a test provides additional cues in written form which materially assist the decoding of the recording.

(20) *++ when I um + because I have no um no um + text, I cannot follow + so I don't know when I have er+ a paper + I can trace and and focus on the key word what the + what the exam want me to do + um + even though I don't I can't get the main idea but + um that's the text er require + you just fill in the key words + but when I do the second test um + you have to follow the speaker line by line+ and you don't know um + what will the key words come*. (L2: 68)

---

(21) why did you find it more difficult?

because er because I I have no paper + I have er some some main ideas from the paper + I read it +er + before I I read the answer I listened the the the cassette yeah + and I can focus my eyes on some + some er special + some some gap yeah + I can sometimes I can guess + I can guess all what's + what they will talk about yeah + and but this + this one the second one I think er + there is no no some background of my + yeah and no some some information to survey before I listen to this yeah + this this lecture + yeah +maybe if you give some papers like this + I can I can finish these very well. (N2: 48)

(22) why did you find it more difficult?

um + if + er if I have this something like this to ask me to do some + do some test er + maybe I can do it, but

so if you have a sheet or an answer sheet it helps you.

yeah yeah + er they can help me to um + to get some key words + and I use the key word to find the answer + but just listen and do some write + I I find I think is more difficult. (R2:45)

Of course, there may be a gap between the respondents' faith in the assistance provided by these cues and what actually occurs in practice. There was evidence in the verbal reports that reliance upon the wording of the task sheet sometimes leads to inefficient strategies which handicap the candidates' ability to extract meaning from the recording. One recurrent problem took the form of a respondent switching attention from the recording to the written text and missing mention of the point that provided the correct answer. Examples of this have already been given in protocol extracts (3) and (4) while (5) shows an instance of a simple match at word level throwing up two alternatives, with no criterion for distinguishing between them if wider context is missing. Problems also arise when candidates lose their way – either overlooking an item on the task sheet (looking for a match for item 35 when the speaker is still covering point 34) or failing to notice a relevant point in the recording (looking for a match for item 35 when the speaker has already moved on to 36).

(23) *um ++ I don't + I don't have that that answer sorry*.

is that because you didn't understand everything or you didn't recognise a word or what?

yeah.

what what was the problem there?

er I can't understand er the recording + it it still talk about air +but the record is about air.

you were looking at the wrong sentences?

yeah yeah.

when the recording was talking about something else.

yeah.

OK + what about number 35?

I just haven't found it. (laughs)

so that was the same thing?

yeah.

looking at sentences at the beginning when you hadn't realised it had + it had moved on?

mhm. (R1: 35)

There is also, as ever, the issue of the limits to the attention capacity which a listener / reader has available. At times, it seemed likely that the participant had lost track of the recording as a result of lending too much attention to possible written cues.

(24) why was that difficult?

um because er + mhm + I I haven't prepared + I haven't warm up to listen + not really ready to listen.

right + you weren't ready + but was that because you were looking at the paper? + or because you don't know the voice of the person who was speaking?

er no.

or you don't know the topic + or what?

um I think I I don't know the + I just look at the paper sometimes.

so you were looking at the paper?

yeah the paper.

so you weren't really listening.

yeah just focus on the the word + probably the speaker might might not say that word er + so I missed it. (S1: 176)

5 DISCUSSION

The study provided a number of useful insights into the way in which candidates respond to two types of test method (gap filling and MCQ) that are quite widely used in IELTS Listening Paper 4. The insights enable us to gauge the extent to which the cognitive processes adopted resemble those that candidates would employ in a real-life lecture-listening situation.

5.1 The use of test-wise strategies

It was apparent from the protocols that the participants had adopted a number of strategies which reflected the nature of the test rather than the demands of lecture-listening or the kinds of gap in understanding that are caused by limited knowledge of L2. The extent and form of these strategies varied quite considerably from one participant to another. In some cases (especially the group from Mainland China), there was evidence of test strategy training, as shown by their use of terms like 'key word'. The training was by no means always beneficial; indeed, it quite often led to a dependence on the written text (itself a challenge for the Chinese learner) which reduced the amount of attention given to the spoken signal.

Firstly, many participants made use of cues provided by the wording of the items.

> Participants reported using a word from the task sheet in order to locate the relevant information in the listening text. Here, they particularly took advantage of collocates (listening for *level* and *frequency* in Text A so as to target the word which preceded them). The location of the correct items was not always accompanied by an understanding of what had been said (witness the two participants who interpreted *low frequency* in terms of infrequency).
>
> Participants used a classic 'key word' strategy, listening out for content words from the task sheet that appeared to be important to the topic or listening out for associates and synonyms of those words.
>
> Participants made use of lists and sequences of words. Where, for example, the gap filling task showed a gap at the end of a list, they listened out for the last word.

Secondly, they used the ordering of items on the worksheet as a kind of checklist with which to approach the recording. Here, they relied on the convention that the order of the questions closely follows the order in which the information occurs in the recording. Several of them also recognised the constraints upon a test setter when designing a gap-filling exercise where a test only permits one hearing of the recording. The information targeted needs to be quite widely spaced to allow participants to tune out partially in order to focus attention on the missing word (and pay due heed to its spelling) before tuning in again to anticipate the next piece of information.

It became evident that using test items in this way to direct the listening process involved a great deal of switching of attention between task sheet and recording. It was also sometimes counter-productive in that it led to participants missing a piece of information when they were consulting written text (see extracts (3) and (4)).

For the test to achieve cognitive validity as a predictor of real-life behaviour, the methods and material used need to replicate at least some of the processes which apply in the special circumstances of academic listening. As we have already noted, a learner in an academic context can certainly expect written input that supports the spoken. It might take two forms: a handout giving an overview of the lecture and/or PowerPoint slides providing visual support for individual points. The critical consideration for cognitive validity lies not in the availability of that input but in how the listener uses it.

The protocols in this study made it clear that much of the use of the written input was at the level of the word or lexical phrase rather than the level of the idea. Instead of extracting a proposition from the test item and then matching it against a proposition expressed by the speaker, the candidate seems typically to use the lexical content of the items to provide cues with which to locate information in the text. The candidate's thinking operates in the direction: *written lexical input – spoken lexical input*.

Compare that with the visual support in a real lecture-listening context. Second language listeners might use the headings in a handout as 'signposts' in order to impose a structure on what is being heard; here there are perhaps parallels with the convention that test items follow the order of information in the recording. They might even attempt some matching at the level of word or lexical phrase like that observed in this study, though without the strategic goals of a test taker. But much of the processing would proceed in the opposite direction: with the listener first picking up a string of words or an idea in the spoken input and then checking it against the handout to confirm that the point in question had been fully understood. A good handout would also be transparent: there would be none of the ambivalence of the multiple-choice item.

As for PowerPoint slides, in a good presentation aimed at a native speaker audience, they tend to operate at propositional level. Whether or not they contain the actual words the lecturer uses, they serve to identify discrete points of information which anticipate or accompany those conveyed by the oral signal. Except in the case of a listener with extremely limited knowledge of L2, they thus provide cues at the level of the idea, not the word.

To be sure, the gap-filling exercise could be said to achieve some simple ecological validity in that it simulates the kinds of notes that a lecture listener might take. The argument is not entirely convincing when the items employed consist of a series of micro-propositions of varying importance without an argument structure to hold them together. But, from a cognitive angle, it is not so much the notes as the gaps which are an issue. Their effect is to fix the candidate's attention at the level of the word or short phrase, giving rise to precisely the types of word-based strategy that have been commented on.

To summarise, while written input is indeed available to support the type of listening that takes place in an academic context, it is unlikely that it would be used in the same way as it is in the test conditions studied. The evidence of these test-wise strategies therefore raises a first set of concerns about the cognitive validity of the methods that were featured.

5.2 Shallow processing in the test condition

A second area of concern follows directly from the first. The protocols suggested that much of the processing was at a very local level. A number of participants who had scored quite well in the test condition were unable to report the two main topics of the lecture in question, to expand upon what the lecturer had said or to trace links between the points that were made. Some showed that they were quite aware of having focused their attention on lexical matches rather than on wider meaning.

(24) *if they um ++ how to say? + what I have to do I have to fill words + so I don't er + listen for the meaning of the whole test text + I am choosing these words + if I have to understand meaning and then write an essay it will be another* (Y1: 40)

Strikingly, participants showed themselves more able to identify main points for both recordings in the non-test condition, which had been hypothesised to be the more demanding task.

There would appear to be at least three reasons for this finding:

the extent to which the test methods and items were dependent upon word matching

the targeting of certain points which were not central to the main argument (e.g. the fact that low frequency noise does not pass through trees, the weight of a large shark)

the cognitive load imposed by the test methods (to be discussed in due course).

An earlier brief characterisation of the meaning construction process in listening suggested that it included the important processes of distinguishing main points from subsidiary ones and of recognising the argument relationships that link propositions. It may indeed be difficult to ensure that these processes feature in any test of L2 listening (important though they are to lecture listening expertise). All one can say here is that they did not seem to have played a significant part when participants in the test condition were asked to report at a global level.

5.3 Distinctive processes in the test and lecture-listening conditions

There was evidence on three counts suggesting a degree of mismatch between the processes demanded by the test and those demanded by a 'free' lecture listening situation. Firstly, no correlation was found between the scores achieved by participants in the administration of the test and the number of micro-propositions reported by them when note-taking and not required to answer specific questions. Secondly, most participants reported differently on the two tasks – expressing the view that one or the other was less demanding. Thirdly, a number of respondents with IELTS and Reading Listening scores at the lower end of the target range performed badly in the test condition but well (in one case extremely well) in the note-taking condition.

The researcher's working hypothesis was that respondents would tend to report the note-taking task as harder than the test-based one, on the grounds that the written items in the test supply the candidate with a schematic framework for the passage that is to be heard. This indeed was what a number of them reported. However, entirely contrary to expectations, nearly a third of participants reported that they found the note-taking task easier than the test. They included both those of European origin and those of Far Eastern origin; both respondents with higher previous test scores and those with lower. Some of them averred that tracking questions made them nervous (extracts (14) to (16)). They specifically mentioned the need to focus on detail in the test, with the accompanying danger that a word or phrase would be overlooked (extracts (15) and (16)).

These reactions would seem to be a consequence in particular of the time-constrained nature of the exercise. Candidates are only allowed to hear the recording once, increasing their fear that they may overlook a low-level detail. They are also sensitive to a phenomenon, for which there is evidence in the protocols, where a listener fails to match an item to the relevant piece of information in the text and goes on listening for it long after it is past – thus missing the answers to subsequent items as well.

The researcher had assumed that the note-taking task would be more cognitively demanding than the test with its accompanying written support. But he had overlooked the important factor of *the additional demands imposed by handling two different sources of information in two different modalities*. They are hinted at in the comments in extract (17) where Participant AA2 expresses concern about monitoring his spelling at the same time as attending to the listening passage. But they emerge most clearly in the following extracts:

(25) *if I don't right now also I don't know if it is correct + and um it is hard to write to read all the tasks before listening + it is better because I when I am filling the first part I don't remember what is following + and when we listen for the next part I have while I'm listening +I have to read and to know what do they want to do*. (Y1: 122)

---

(26) do you um... + do you manage to read and write and listen OK when you...

no no + this time I'm not manage this good.

mhm.

and . . .

is that usual?

Yes + that's usual.

So you have problems with reading writing and listening at the same time + but when you're doing note-taking you're writing and listening.

yes + I think er writing um + quickly we would be happy to er + memorise the lecture.

Mhm + so you think that it's OK to write and listen...

yes yes.

but you find it difficult to read and write and listen?

yes yes. (Z2: 110)

What the researcher had not allowed for – and what emerged in these and some of the other protocols – was the complexity of the tasks demanded by the two test methods represented here. Gap-filling might appear to be an activity that closely approximates to the type of note-taking that takes place in a lecture. But it does not really do so, because the notes have not been generated by the candidate and therefore represent an unseen text that has to be mastered. The test format demands a combination of reading, listening and writing. Attention needs to be switched between the three skills (with the added complication of Cambridge ESOL's accurate spelling requirement) and even at times divided between them. As already noted, human attention is limited in capacity and attention dividing activities make complex demands upon the processor. Something similar can be said of MCQ. It has often been remarked that MCQs load heavily on to the reading skill because of their complexity. But the issue here is not so much the part played by reading as, once again, the requirement upon the candidate to manipulate two skills, both demanding high levels of attention. Wickens' multiple resource theory (1984) suggests there may be particular tensions when two sources of information share a single channel, as the two receptive skills would be deemed to do by some commentators.

In this respect then, the test methods used in connection with Paper 4 appear to make considerably *heavier cognitive demands* upon the candidate than would a real-life situation.

5.4 Additional cognitive demands of note-taking

That said, there was incidental evidence that in certain other areas the note-taking task was more demanding for participants than undertaking the test. Participants showed themselves to be vulnerable in three areas in particular when performing in lecture conditions.

a Constructing meaning representations

Without the support of the kind of outline that is provided by a set of test items, participants were much more prone to construct their own hypotheses as to the main direction of the speaker's argument or the main themes of the lecture. These hypotheses could be close to the truth but they could also lead the listener into establishing meaning representations which did not accurately represent what was in the recording. In forming their assumptions, participants were assisted or misled by their knowledge of the topic (particularly so with the shark text) and sometimes by their intuition as to what might be a current angle on the topic (protecting trees, protecting sharks).

It has to be said, though, that mistaken hypotheses were by no means restricted to the note-taking condition; they were also observed in the test condition despite the availability of supportive written text. What seemed to be more prevalent among note-takers was a tendency to construct an elaborate meaning representation on the basis of a single word – sometimes a word that had not been correctly recognised. Thus, three participants reported on shark *machines* (= 'meshing') while one misheard the word *beach as breed* and interpreted the entire lecture as being about the propagation of shark species.

b propositional density and complexity

Without targeted questions, participants seemed prone to lose their way when confronted with sequences which were particularly dense propositionally or complex in terms of the relationships between the propositions. An example of the first was that several of them commented on the heavy factual load of the shark lecture. An example of the second was that very few of them managed to make sense of the exposition of how high buildings created wind tunnels.

c lack of selectivity

Some participants had difficulty in distinguishing central facts from peripheral ones when reporting orally on Text B (see 'peripheral' in Tables 2 and 3).

In these three areas, the note-taking task was arguably more demanding. The point at issue is that, here again, there would appear to be a lack of fit between the demands of the test formats and those of the target behaviour. A key to handling the types of issue that have been identified lies in the listener's ability to *self-monitor*, checking the relevance and reliability of incoming information in the light of the meaning representation built up so far. This aspect of lecture listening is sidelined when the listeners have detailed written prompts that help build a representation for them, regardless of what they have extracted from the recording.

6 RECOMMENDATIONS

6.1 Some tentative suggestions for IELTS testing in this area

It should be stressed at this point that the view of cognitive validity presented in the report is a somewhat idealised one. It is clearly not possible for any test to replicate all the processes that a real-life listening event demands. In addition, exam boards have to observe a number of important considerations – not least, the need to achieve marker reliability. Any proposals that are made in this section must therefore remain tentative and subject to the usual constraints associated with efficient test administration.

Nevertheless, the study has served to highlight several ways in which current test formats are either more cognitively demanding than a lecture-listening task or fail to embrace some of its more important aspects (selecting relevant information, linking points made by the speaker, building a macro-/micro- comprehension structure, self-monitoring). It should not be impossible to adjust or replace the methods that are used in the IELTS listening section 4 in order to make this test a more sensitive detector of the ability to perform in real-life academic listening contexts. Some suggestions follow.

6.1.1 Test method

The gap filling and MCQ formats as they are currently employed are worth reconsidering. They appear to make cognitive demands upon the candidate which exceed those of normal lecture listening. The former has the unfortunate effect of focusing candidate attention at word level and providing *gratis* a great deal of the structure of the lecture which it should be the listener's responsibility to construct. The latter imposes heavy reading demands. Both foster a practice of switching attention away from the recording to the written modality (seen by learners as easier to process because it can be consulted over and over again).

Ways of refining the gap filling format might be

to focus more strictly upon points which are central to the main argument

to target propositions at macro- as well as at micro- level – perhaps by featuring two short sets of notes of which one provides an overview of the lecture

to rely more heavily upon paraphrase than at present so as to avoid word-matching strategies

to provide a skeleton outline of the lecture rather than simulated notes, with macro- as well as micro-elements to be filled in

(given the number of correct answers in the data which would have been disallowed by the strict marking scheme) to allow more latitude both on acceptable responses and on spelling.

In many ways, however, it would be advisable to abandon this format, given its heavy cognitive demands and the way it fosters test-wise strategies. More valid alternatives would require the candidate to write a summary of the lecture or to insert notes under various headings (not necessarily following the order of the text). However, these methods would certainly create problems of marker reliability. More practical alternatives might include

jumbled propositions (paraphrased from the recording) for the candidate to fit into a skeleton outline of the lecture

a complete and coherent paraphrased summary of the text with gaps for candidates to fill

(to test structure building) a paraphrased summary of the text with gaps for candidates to insert connectives chosen from a limited set.

[The first two of these would need to be carefully controlled to ensure that they did not load too heavily on to reading.]

Ways of adapting the MCQ format would be:

to focus more strictly upon points which are central to the main argument

to provide shorter options and options which are less finely differentiated so as to reduce the reading load.

A rather threadbare argument in favour of MCQ is that it replicates what is in the mind of a listener, who approaches a lecture with expectations that need to be tested. This does not hold up from a process perspective in that accessing those expectations requires a complex reading operation. A more viable alternative along these lines might be to expand the use of the traditional 'true / false / not mentioned' format in section 4. Even better would be to ask a candidate to read a complete and coherent (but concise) summary of the lecture which was incorrect in some respects and to underline the propositions which were wrong.

The most ambitious but also the most cognitively valid alternative would be to ask candidates to listen to not one but two lectures on the same topic and to collate the information from them into a table.

6.1.2 Multiple play

There are a number of reasons for the present policy of only allowing one hearing of the text. (for a rationale, see Geranpayeh and Taylor 2008, p 4). One is historical: the single-play stipulation has always set IELTS apart from the exams of the main Cambridge suite. One is practical: double play extends the length of listening time and thus potentially restricts the length, number and variety of the recordings that can be employed within the time frame of the test. However, it would appear that the convention has a number of unfortunate side effects. As evidenced in this study, it creates tension in the candidate who is afraid of missing a point (often a point of detail) and it fosters test-wise strategies at the expense of overall meaning. In other words, it exercises an effect upon the cognitive processing that takes place in the course of the test.

An 'ecological' argument is sometimes put forward that in real life lecture listeners only hear a point once and have to grasp it or lose it; but it is not entirely convincing in the context of a test and moreover one that is based upon audio input. Firstly, a real-life lecture has far greater redundancy than the type of brief recording that, for obvious practical reasons, features in an international exam. The lecture mode relies quite heavily upon rephrasing and repetition to underline critical points; it also has a distinctive discourse structure in which the lecturer provides an outline at the outset and a summary at the end. Candidates hearing a short IELTS recording do not have the benefit of these features; small wonder that the one-off opportunity to grasp a point sometimes contributes to the kind of stress mentioned in the protocols. In addition, the candidate who hears an audio recording of a lecture cannot be said to be in a situation that resembles a real-life one in terms in cognitive

terms. Processing demands are affected by the fact that the candidate has no access to Powerpoint support of the kind that would normally be available or to the paralinguistic cues that would normally be provided by the lecturer.

Also persuasive is the evidence of what listeners do when they know that they will hear an audio recording twice. As Buck (1990) testifies, they tend to listen at a rather local level during the first play; during the second, they engage in structure building, assembling the points they have identified into a coherent whole and recognising the logical connections between them. It was precisely this element that was found to be absent in the accounts of many of the participants in the test-taking condition. They proved capable of scoring IELTS points by providing the locally-based information that the tests required; but they were not able to achieve what successful lecture attendance would normally demand – a coherent account of the main points of the lecture and the ways in which they were linked. The convention of only allowing a single play would thus seem to be implicated in the low level of processing in which candidates engaged. It also contributed importantly to the heavy cognitive demands imposed by the gap filling task in that it required candidates not only to operate in three different ways (reading, listening and writing) but to do so under extreme pressures of time and attention allocation, given that they were unable to listen again to check their impressions.

Whatever the ecological arguments (and it has been suggested that they are not strong), the present study seems to show that the single play stipulation detracts from cognitive validity. The IELTS partners might perhaps consider the benefits of a double play.

6.1.3 Propositional density and complexity

The comments of a number of participants about the texts they heard (as against the tasks they performed) indicate a level of concern with parts of the recording that were dense in terms of the amount of detail they contained or complex in terms of the links between propositions. These considerations should perhaps be accorded greater weight by test setters. A transcript that suggests that a recording is rich in details that can be tested may seem to be an attractive proposition but may make unfair cognitive demands of the candidate – not least because of the point made in the previous section that candidates only hear a short presentation and cannot benefit from the more elaborate discourse structure and the level of redundancy that counter-balance informationally dense sections in a normal lecture context.

6.1.4 Greater authenticity

Finally, it is worth recording that a real-life lecture is a multi-modal event to which a number of sources of information contribute. Many of them are absent in the current format, reshaping the cognitive operations that are required of the listener. They include:

handout material

Powerpoint slides

facial expression and gestures of the lecturer

the tendency of the lecture mode towards redundancy in the form of repetition and rephrasing

Long-term, it would be desirable to ensure that the IELTS test (and particularly the lecture-listening component) approximates more closely to these real-life conditions. That would entail taking advantage of current technology to ensure that the input to the candidate has visual as well as auditory components and that the components replicate as closely as possible those available to the academic listener. Clearly, full account would need to be taken of the limited technological resources in some parts of the world where the test is taken; this might well delay the use of DVD or downloadable materials. However, innovation is likely to prove necessary at some stage if the test is to increase its validity as a predictor of actual lecture-listening behaviour.

6.2 Limitations of the study and further research

The most suitable way of obtaining the evidence needed for this study was felt to be by retrospective verbal report. The method is demanding in terms of time and the type of analysis involved; and only permits the study of a relatively small sample population. Its findings therefore need to be accompanied by the rider that they can only be indicative. It would certainly be of value to extend the study by examining the test-taking and lecture-listening behaviour of a further group of participants.

It would also be valuable to extend it by using the same methodology but employing other past IELTS papers. This might enable one to establish the extent to which characteristics of the recording or of the test method are factors in the types of process that candidates are likely to adopt.

Attempts were made to balance that population across first languages. Nevertheless, the size of the study did not permit of any detailed investigation of the possible effects upon cognitive processes of a) first language, b) cultural and educational background, or c) preparation in the home country for IELTS. All of these factors merit further exploration – possibly in a limited set of country-by-country studies.

The issue of cognitive validity seems likely to gain in importance as a consideration in test design. What will surely be needed long-term are longitudinal studies which attempt to evaluate the predictive power of an IELTS Listening score. These might track former IELTS candidates during their first year at an English-medium university. Ideally, one could video-record live lectures within their discipline and re-run them to the participants in order to assess at intervals their developing ability to process the content. A study of this kind should certainly make use of the type of verbal report that has been employed here; it would be instructive to see if participants' strategies changed as they gained more experience of lecture listening and better knowledge of L2.

That said, listening development is a complex area to which many different factors contribute. Quite apart from the very varied ways in which individuals respond to the challenge of L2 listening, there are considerations such as distance of L2 from L1, familiarity with western patterns of logic, extent of integration into the host community, motivation, grasp of the discipline being studied and the communicative imperative felt by the listener. All this suggests that any longitudinal research will need to rely upon a whole series of case studies. There seems to be scope for a great deal of investigation in this area in years to come.

ACKNOWLEDGEMENTS

I am extremely grateful to Ros Richards, Director of the Centre for Applied Language Studies at Reading University for allowing me access to students and facilities at the centre. I am also very grateful to the unfailingly supportive staff of CALS (especially Colin Campbell, Jonathan Smith and John Slaght) who allowed me to contact their students and assisted with evidence on the students' backgrounds. Particular mention should be made of the technical advice I received from Pete Cox of Reading University and from Mark Huckvale of University College London.

Special thanks are owed to the students from many parts of the world who participated in the data collection: for their interest in the project and for the engaged and helpful way in which they reported on their experience of undertaking the tasks.

Many thanks to Professor Cyril Weir of the University of Bedfordshire for some stimulating conversations on the topic of cognitive validity.

Finally, I express my appreciation to the British Council for funding what I believe to be much-needed research into the extent to which the processes underlying test performance replicate the processes that would be applied in a non-test context. I trust that it will be of assistance to future test design.

REFERENCES

Alderson, C, 2000, *Assessing reading*, Cambridge University Press, Cambridge

Bachman, L, 1990, *Fundamental considerations in language testing*, Oxford University Press, Oxford

Baxter, G P, and Glaser, R, 1998, 'Investigating the cognitive complexity of science assessments', *Educational Measurement: Issues and Practice*, vol 17, no 3, pp 37-45

Brown, G, 1995, *Listeners, speakers and communication*, Cambridge University Press, Cambridge

Brown, J D, and Rodgers, T, 2002, *Doing second language research*, Oxford University Press, Oxford

Buck, G, 1990, *The testing of second language listening comprehension*, Unpublished PhD thesis, University of Lancaster

Buck, G, 2001, *Assessing listening*, Cambridge University Press, Cambridge

Cambridge ESOL, 2005, *Cambridge IELTS 4*, Cambridge University Press, Cambridge

Clapham, C, 1996, *Studies in language testing 4: the development of IELTS*, UCLES / Cambridge University Press, Cambridge,

Cohen, A, 1998, *Strategies in learning and using a second language*, Harlow, Longman

Dunkel, P, Henning, G, and Chaudron, C, 1993, 'The assessment of an L2 listening comprehension construct: A tentative model for test specification and development', *Modern Language Journal*, vol 77, pp 180-191

Ericsson, K A, and Simon, H A, 1993, *Protocol analysis: verbal reports on data*, 2nd ed, MIT Press, Cambridge MA

Faerch, C, and Kasper, G, 1987, *Introspection in second language acquisition research, Multilingual Matters*, Clevedon

Field, J, 2004a, *Psycholinguistics: the key concepts*, Routledge, London

Field, J, 2004b, 'An insight into listeners' problems: too much bottom-up or too much top-down?' *System*, vol 32, pp 363-377

Field, J, 2008a, *Listening in the language classroom*, Cambridge University Press, Cambridge

Field, J, 2008b, 'The L2 listener: type or individual?', *RCEAL Working Papers in English and Applied Linguistics*, vol 12, pp 13-32

Field, J, forthcoming, 'Cognitive validity', in *Examining speaking*, ed L Taylor, Cambridge University Press, Cambridge

Gaskell, G (ed), 2007, *The Oxford handbook of psycholinguistics*, Oxford University Press, Oxford

Gass, S M, and Mackey, A, 2000, *Stimulated recall methodology in second language research*, Erlbaum, Mahwah, NJ

Geranpayeh, A, and Taylor, L, (2008) 'Examining listening developments and issues in assessing second language listening', in *Cambridge ESOL*, Research notes, 32, pp 2-5

Gernsbacher, M A, 1990, *Language comprehension as structure building*, Erlbaum, Hillsdale, NJ

Glaser, R, 1991, 'Expertise and assessment', in *Testing and cognition*, eds M C Wittrock and E L Baker, Prentice Hall, Englewood Cliffs, pp 17-30

Kellogg, R, 1995, *Cognitive psychology*, Sage, London

Long, D R, 1990, 'What you don't know can't help you', *Studies in Second Language Acquisition*, vol 12, pp 65-80

Lynch, T, 1994, 'Training lecturers for international audiences', in *Academic listening: Research perspectives, ed J Flowerdew*, Cambridge University Press, Cambridge

Lynch, T, 2004, *Study listening*, Cambridge University Press, Cambridge

McDonough, J and McDonough, S, 1997, *Research methods for English Language teachers*, Arnold, London

Oxford, R, 1990, *Language learning strategies: what every teacher should know*, Newbury House, Rowley, MA

Richards, J C, (1983) 'Listening comprehension: Approach, design, procedure', *TESOL Quarterly* vol 17, no 2, pp 219-39

Rost, M, 1990, *Listening in language learning*, Longman, Harlow

Schmidt-Rinehart, B, 1994, 'The effects of topic familiarity on second language listening comprehension', *Modern Language Journal*, vol 78, no 2, pp 179-189

Shaw, S, and Weir, C, 2007, *Examining writing*, Cambridge University Press, Cambridge

Styles, E, 2006, *The psychology of attention*, 2nd ed, Psychology Press, Hove

Tulving, E, 1983, *Elements of episodic memory*, Oxford University Press, New York

Van Dijk, T A, and Kintsch, W, 1983, *Strategies of discourse comprehension*, Academic Press, New York

Vandergrift, L, 2005, 'Relationships among motivation orientations, metacognitive awareness and proficiency in L2 listening', *Applied Linguistics* vol 26, pp 70–89

Vandergrift, L, Goh, C, Mareschal, C, Tafaghodatari, M H, 2006, 'The Metacognitive Awareness Listening Questionnaire (MALQ): Development and validation', *Language Learning*, vol 56, pp 431–462

Weir, C, 2005, *Language testing and validation: an evidence-based approach*, Palgrave Macmillan, Basingstoke

Wickens, C, 1984, 'Processing resources in attention', in *Varieties of attention* eds R Parsuraman and D R Davies, Academic Press, Orlando, FL

APPENDIX 1: RECORDED TEXTS USED IN THE STUDY

TEXT A

[Test 1, Section 4, *Cambridge IELTS with Answers*, 4, 2005, pp 134-5]

Good day, ladies and gentlemen. I have been asked today to talk to you about the urban landscape. There are two major areas that I will focus on in my talk: how vegetation can have a significant effect on urban climate, and how we can better plan our cities using trees to provide a more comfortable environment for us to live in.

Trees can have a significant impact on our cities. They can make a city, as a whole, a bit less windy or a bit more windy, if that's what you want. They can make it a bit cooler if it's a hot summer day in an Australian city or they can make it a bit more humid if it's a dry inland city. On the local scale – that is, in particular areas within the city – trees can make the local area more shady, cooler, more humid and much less windy. In fact trees and planting of various kinds can be used to make city streets actually less dangerous in particular areas. How do trees do all that you ask?

PAUSE INSERTED

Well, the main difference between a tree and a building is a tree has got an internal mechanism to keep the temperature regulated. It evaporates water through its leaves and that means that the temperature of the leaves is never very far from our own body temperature. The temperature of a building surface on a hot sunny day can easily be twenty degrees more than our temperature,. Trees, on the other hand, remain cooler than buildings because they sweat. This means that they can humidify the air and cool it – a property which can be exploited to improve the local climate.

Trees can also help to break the force of winds. The reason that high buildings make it windier at ground level is that, as the wind gets higher and higher, it goes faster and faster. When the wind hits the building, it has to go somewhere. Some of it goes over the top and some goes around the sides of the building , forcing those high level winds down to ground level. That doesn't happen when you have trees. Trees filter the wind and considerably reduce it, preventing those very large strong gusts that you so often find around tall buildings.

PAUSE INSERTED

Another problem in built-up areas is that traffic noise is intensified by tall buildings. By planting a belt of trees at the side of the road, you can make things a little quieter, but much of the vehicle noise still goes through the trees. Trees can also help reduce the amount of noise in the surroundings, although the effect is not as large as people like to think. Low frequency noise, in particular, just goes through the trees as though they aren't there.

Although trees can significantly improve the local climate, they do however take up a lot of space. There are root systems to consider and branches blocking windows and so on. It may therefore be difficult to fit trees into the local landscape. There is not a great deal you can do if you have what we call a street canyon – a whole set of high-rises enclosed in a narrow street. Trees need water to grow. They also need some sunlight to grow and you need room to put them. If you have the chance of knocking buildings down and replacing them, then suddenly you can start looking at different ways to design the streets and to introduce . . . (fade out)

TEXT B

[Test 4, Section 4, *Cambridge IELTS with Answers*, 4, 2005, p 151]

Today we're going to look at one of my favourite fish – the shark. As you know, sharks have a reputation for being very dangerous creatures capable of injuring or killing humans, and I'd like to talk about sharks in Australia.

Sharks are rather large fish, often growing to over ten metres, and the longest sharks caught in Australia have reached sixteen metres. Sharks vary in weight with size and breed, of course, but the heaviest shark caught in Australia was a White Pointer – that weighed seven hundred and ninety-five kilograms – quite a size! Sharks have a different structure to most fish: instead of a skeleton made of bone they have a tough elastic skeleton of cartilage. Unlike bone, this firm, pliable material is rather like your nose, and allows the shark to bend easily as it swims. The shark's skin isn't covered with scales, like other fish: instead, the skin's covered with barbs, giving it a rough texture like sandpaper. As you know, sharks are very quick swimmers. This is made possible by their fins, one set at the side and another set underneath the body, and the tail also helps the shark move forward quickly.

Unlike other fish, sharks have to keep swimming if they want to stay at a particular depth, and they rarely swim at the surface. Mostly, they swim at the bottom of the ocean, scavenging and picking up food that's lying on the ocean floor. While most other animals, including fish, hunt their prey by means of their eyesight, sharks hunt essentially by smell. They have a very acute sense of smell – and can sense the presence of food long before they can see it.

PAUSE INSERTED

In Australia, where people spend a lot of time at the beach, the government has realised that it must prevent sharks from swimming near its beaches. As a result, they've introduced a beach-netting programme. Beach-netting, or meshing, involves setting large nets parallel to the shore: this means that the nets on New South Wales beaches are set on one day and then lifted and taken out to sea on the next day. When shark meshing first began, in 1939, only the Sydney metropolitan beaches were meshed – these beaches were chosen because beaches near the city are usually the most crowded with swimmers. Ten years later, in 1949, systematic meshing was extended to include the beaches to the south of Sydney. As a result of the general success of the programme in Sydney, shark-meshing was introduced to the state of Queensland around 1970. The New Zealand authorities also looked at it, but considered meshing uneconomical – as did Tahiti in the Pacific. At around the same time, South Africa introduced meshing to some of its most popular swimming beaches.

When meshing began, approximately fifteen hundred sharks were caught in the first year. However, this declined in the years that followed, and since that time, the average annual catch has been only about a hundred and fifty a year. The majority of sharks are caught during the warmest months, from November to February, when sharks are most active and when both the air and ocean are at their maximum temperature.

PAUSE INSERTED

Despite quite large catches, some people believe that shark meshing is not the best way to catch sharks. It's not that they think sharks are afraid of nets, or because they eat holes in them, because neither of these is true. But meshing does appear to be less effective than some other methods, especially when there are big seas with high rolling waves and strong currents and anything that lets the sand move – the sand that's holding the nets down. When this moves, the nets will also become less effective.

APPENDIX 2: TASKS USED IN THE STUDY

TASK A

Section 4
Questions 31–40

Complete the notes below

Write **NO MORE THAN TWO WORDS** for each answer.

The urban landscape

Two areas of focus:

the effect of vegetation on the urban climate

ways of planning our 31 ..... better

Large-scale impact of trees:

they can make cities more or less 32 .....

in summer they can make cities cooler

they can make inland cities more 33 .....

Local impact of trees:

they can make local areas

– more 34 .....

– cooler

– more humid

– less windy

– less 35 .....

Comparing trees and buildings

Temperature regulation:

trees evaporate water through their 36 .....

building surfaces may reach high temperatures

Wind force:

tall buildings cause more wind at 37 ..... level

trees 38 ..... the wind force

Noise:

trees have a small effect on traffic noise

39 ..... frequency noise passes through trees

Important points to consider:

trees require a lot of sunlight, water and 40 ..... to grow

TASK B

Section 4
Questions 31–40

Complete the notes below

Write **NO MORE THAN THREE WORDS AND/OR A NUMBER** for each answer.

Sharks in Australia

Length	largest caught:	16 metres
Weight	heaviest	31 ..... kg
Skeleton	cartilage	
Skin texture	rough barbs	
Swimming aids	fins and 32	
Food	gathered from the ocean 33	
	sharks locate food by using their 34	

Questions 35–38

Choose the correct letter A, B or C

35 Shark meshing uses nets laid
- **A** along the coastline
- **B** at an angle to the beach
- **C** from the beach to the sea.

36 Other places that have taken up shark meshing include
- **A** South Africa
- **B** New Zealand
- **C** Tahiti.

37 The avarage number of sharks caught in nets each year is
- **A** 15
- **B** 150
- **C** 1,500

38 Most sharks are caught in
- **A** spring
- **B** summer
- **C** winter

Questions 35–38

Choose two letters A–G

Which two factors reduces the benefits of shark nets?
- **A** nets wrongly positioned
- **B** strong waves and currents
- **C** too many fish
- **D** sharks eat holes in nets
- **E** moving sands
- **F** nets too short
- **G** holes in nets scare sharks

APPENDIX 3: SAMPLE TRANSCRIPTIONS: PARTICIPANT R

TEXT A (TEST CONDITION)

Italics indicate researcher's turns

1 *right + 31?*
2 er 31 ways to plan our cities' trees better.
3 *sorry, what was the answer?*
4 'city'.
5 *'city' OK + why did you choose the word 'city'?*
6 um ++ I can hear the sentence.
7 *so you heard someone talking about cities?*
8 yeah.
9 *that's why you put 'city' in + OK fine + um number 32?*
10 er + sorry + I missed er the answer + the answer I think is wrong + I write here.
11 *OK that's OK + so you didn't hear something that would give you the answer?*
12 um I hear + I hear the er the sentence clearly but I I lost er + the the title + er + I
13 I missed the title + the recording is is er faster than I thought +
14 *so you heard the sentence but you didn't understand all the words?*
15 no I understand all the words but I I didn't er record down on the paper + I I
16 clearly know the the meaning of the recording.
17 *Yeah but you didn't why didn't you write it on the paper?*
18 Mhm + er ++ I + er ++ er + sorry I er the time I just heard + er the the time gave
19 me to look out all that's all + the test is is short so I didn't think I'd finish all the
20 all the title so I + I missed the the key words.
21 *so you missed it because you were reading the sentence+*
22 Yeah.
23 *when when you were listening.*
24 yeah I just find the key words of the test.
25 *OK so what do you think the answer was for 32?*
26 32 ++ um 'comfortable' maybe.
27 *OK right + um right would you give me the answer for 33?*
28 um 'humid'.
29 *yeah + and why did you give that answer?*
30 mhm.
31 *what did you hear the speaker say?*
32 er the speaker say if er ++ plan er um plan more trees come makes inland city
33 more humid + I just caught the end + sentence.
34 *OK great + Um 34?*
35 um ++ I don't I don't have that that answer sorry.
36 *is that because you didn't understand everything or you didn't recognise a word,*
37 *or what?*
38 yeah.
39 *what, what was the problem there?*

40 er I can't understand er the recording + it it still talk about air but the record is
41 about air.
42 *you were looking at the wrong sentence*
43 yeah yeah.
44 *when the recording was talking about something else.*
45 yeah.
46 *OK. What about number 35?*
47 I just haven't found it. (laughs)
48 *so that was the same thing?*
49 yeah.
50 *looking at sentences at the beginning when you hadn't realised that it had it had*
51 *moved on?*
52 mhm.
53 *OK thanks for that + so now we're going to hear a little bit more, yeah.*
54 er again play again?
55 *no no.*
56 go on? OK.

---

57 *OK + have you answered some more? + right um + would you like to carry on?*
58 yeah yeah ++ er just give you the answer?
59 *yeah + 36.*
60 'leaves'
61 *'leaves' + why did you say 'leaves'?*
62 um ++ I got hear the sentence + the er this recording the second recording I I
63 thought I got be some um + I used to the recordings + and the + the voice, so I +
64 *easier.*
65 easier + yeah.
66 *and did you hear her say something about leaves?*
67 um.
68 *yeah + what did she say about leaves?*
69 er can ++ the building surface make +
70 *what did she say? + do you remember what she said about leaves?*
71 just this part this sentence.
72 *right + so the same thing.*
73 yeah yeah.
74 *OK + um now+ what about 37?*
75 37 + um + I think maybe er maybe thirteen, thirty-eight level, or high level. I'm
76 not sure about this answer, but I...
77 *it's a high level +, or what was the other one?*
78 er + er + I just listen er 30 30 30 what I missed it.
79 *mhm.*
80 but the 38 is 'break'

81 *no hang on + you think it's um + you think it's a high level?*
82 yeah.
83 *why did you choose 'high'?*
84 um ++ er I hear I heard the the wind er+ the wind go through the building can get
85 faster and faster er + particularly er in the tall buildings.
86 *mhm.*
87 so so + can cause more wind at a high level.
88 *OK so you heard the word 'high' did you?*
89 yeah.
90 *yeah + OK + um 38?*
91 38 + er 'trees break the wind force' + and er sentence is same as this the the
92 recording sentence.
93 *so you said the trees break the force of the wind?*
94 yeah.
95 *OK + so you actually heard that sentence?*
96 yeah.

---

97 *great + what about 39 and 40?*
98 'low' focusly for for frequently.
99 *low + low frequency?*
100 frequency ah.
101 *right right why did you say that?*
102 er I got the frequenc the frequency that word + because the word in this title is
104 unique + so I just hear er I just look er + where I listen the recording I look for the
105 frequency.
106 *frequency and you heard the word that came before it.*
107 yeah.
108 *OK great + and number 40?*
109 number 40 'room'.
110 *right.*
111 and he er + some the recording give some some interrupt er + because er + he she said
112 she water 'water' before 'the sunlight' + but at end is the room
113 *what does 'room' mean? Do you know?*
114 um space.
115 *right + great + well done ++ OK, I've got two questions about this then + and the first*
116 *one is were there parts of the recording that you found difficult to listen to?*
117 er you mean test + the +
118 *parts of the recording that you found difficult to understand.*
119 um I think all the all all the test all the test is not difficult.
120 *not the task the recording.*
121 rhe recording is not difficult to me.
122 *right.*

123 but bec because I missed the this this this answer + because I have lot of time to to
124 *because you're looking at the text yeah? + because you were reading.*
125 yeah yeah I + have a lot of a long time to to + have nice projects about about
126 listening um + especially this er professional listening so so I missed if give
127 some one some er just one week's I think I can get er better get better.
128 *now my um other question is what do you think the text is about? What were the*
129 *main points that the speaker was making?*
130 um the test was a recording + the recording talk about the er ++ talk about er the
131 tree grow tree in the city and er + I'm not sure the word + er some some some
132 good thing for grow tree in cities er can + can some in spite of the the weather +
133 not the weather the environment and the noise and the temperature and and +
134 some some very important to human lives.
135 *mhm and the part at the end about trees + what did that say?*
136 pardon?
137 *there was something at the end the last part about trees + what did that say?*
138 and the (inaudible) is to grow trees.
139 *mhm.*

TEXT B (NON-TEST CONDITION)

Italics indicate researcher's turns

1 OK um um the shark is actor's favourite fish + er it's it's very long + very long er
2 ten yard fish + they are they really kill humans + and the larger is the shark in
3 Australia named er + 'white + white shark' + and can er + sixteen metres long and er +
4 more than nine hundred kilo kilogram er + and the shark um can swim very quick +
5 and so they have good smelling to help them to find food + er it produce the smell
6 of bloody.
7 *mhm.*
8 at the beginning to er + the beginning maybe the shark attacked human in in 1939 +
9 1939 in beach + 19 + er some beach near the city + and then ten years later in 1949
10 in Sydney + and in 1970 in Queensland + and the shark eat the the food sign food
11 chain + food chain and er ++
12 *what you think the shark + you think that the speaker said that the shark + eats*
13 *everything that is smaller than the shark?*
14 um I I + think the meaning is um + like like er leo in the land + or like human in the
15 war is er + top list of eaten.
16 *so it's the top fish + so it eats all the fish that are smaller?*
17 yeah.
18 *OK.*
19 um.
20 *anything else that they said about sharks?*
21 yeah er + um in in Australia there are some popular beach er to let people to to
22 play er + without er shark attack them + and the shark only live in the er warm

water um +
mhm.
I just got there.
OK so why don't the sharks attack people on these beaches? + do you know?
um + maybe they think that people interrupt them. (laughs)
oh they're frightened of people +
yeah.
is that what you're saying? + on those beaches?
yeah.
so um what do you think the lecture was really about? + what were the important points that the lecturer was making?
um ++ I I don't think there there are some main idea in the + in the recording + and the actor just er described er the shark + the kind of shark in Australia + and er tell told some some + some truths er + for shark attack people um + but the people people for the + of the recording um + don't have the I don't think it have a main body.
OK um was there any part of the recording that you found very difficult to understand? + the beginning or in the middle or at the end?
the end.
the end you found difficult to understand?
yeah.
um did you find this more difficult than the last one the one about the urban landscape?
yeah.
why did you find it more difficult?
um + if + er if I have this + something like this to ask me to do + some do some test
er maybe I can do it but +
so if you have a sheet or an answer sheet it helps you.
yeah yeah + er they can help me to um + to get some key words and I use the key word to find the answer + but just listen and do some write + I I find I think is more difficult.

APPENDIX 4: SAMPLE TRANSCRIPTIONS: PARTICIPANT V

TEXT B (TEST CONDITION)

Italics indicate researcher's turns

so first er 31 is seven hundred and ninety-five kilos + I've but I chose it because
I heard the the number + then +
did you hear anything else + like the word 'kilo' or 'kilogram' or something like that?
I don't remember + I was really focusing on the number (laughs)
but it seemed to be to do with weight?
yeah.
yeah yeah + OK um + um + the second answer the 32 was 'tails' + um + I

heard this word because er + the the speaker said that + sharks um + swimming
with fin and tails and he I don't remember really why but I heard the word +
then...
you understood 'fins' and 'tails'.
yeah yeah + then the the next one 33 I heard 'ocean floor' + and er I knew he
was talking about food the speaker +
you heard 'ocean' and 'floor' together?
yeah.
had you ever heard the word 'floor' with the word 'ocean' before?
No no.
so you identified two separate words together.
yeah.
great.
um and the last one um + the 34 they used er the smell to find some to locate
some food + and well this this point I knew that they had a really good smell + so
er when I heard 'smell' I was sure it was the the answer.

OK 35.
yes OK + so first one the 35 + I'm not sure because er I didn't catch the um the
sentence + but I guess it's 'along the coastline' +but this +
why did you guess that?
because this is my this is my guess (laughs) + I mean um +
did it seem +
the second one + the second one is not logical at all + 'at an angle to the beach' +
it's quite strange + the second one 'from the beach to the sea' I don't really
understand how + how they could put a net from the beach to the sea + so 'along
the coastline' seems +
that would be logical because you understood +
yeah.
it was to do with nets?
yeah yeah + yes + um the the second one the 36 was a bit tricky because er he
mentioned all these places um + but I think it's South South Africa because um +
actually I know this but er + I know they use some some nets but I wasn't sure +
um +
so basically you used your own knowledge?
yeah. (laughs)
any other reason for preferring South Africa?
er because I'm um + I am a surfer + and I know that they are um + there are a lot of
sharks in South Africa + and lot lots of um um + problems with with sharks and
South Africa + um the the answer for the 37 was 'one thousand and fifth
hundreds er + sharks caught' because I heard the answer.
did you hear any other numbers at all?

um I don't think so no.
you just heard '1,500'?
yes I guess + and the last one the 38 was the the + hottest er period + so it's the
summer even if it's November in the south part + it's it's the summer.

39.
OK so the 39 is B because um er + it reduce the benefits of shark nets when the the
+ waves and the currents are strong.
did you hear them say something similar?
um um not really + I heard about the strong waves and currents.
so you heard him use the word 'waves' +
yeah.
and 'currents'.
yeah + I heard these words.
you didn't actually understand what he was saying about them + but...
well um + not really facts er yeah.
logic again.
excuse me?
logic.
yeah and um the second one the + the answer 40 was E 'moving sands'.
why?
because um because I I guess again + but because I heard 'move moving
moving sands' + but I guess it's + it should be hard to fix some nets in the ground
um + which is too soft when the + when the the sand is moving yeah.
*OK thanks + now could you tell me what do you think the main topic or topics of
this um lecture is?*
the main topics is about er + how to keep er sharks away from the beach + to
avoid er injuries and accidents and +
*OK was there any part of the recording that you found difficult? + the beginning
or the middle or the end?*
the middle + um the points um 36 was quite + quite difficult yeah.
why?
because it's tricky when you have to find something but er the speaker men
mentions everything + it's really hard.
for all those things?
yeah + because you don't know if it's true if + or if it's not and yeah it's quite
hard.

TEXT B (NON-TEST CONDITION)

Italics indicate researcher's turns

OK so this lecture's focused on the urban landscape and especially on one point + was about the trees and + and see if trees er could provide any advantages in er urban landscape in a urban area + and it revealed that it's er + very interesting to have trees in cities + because the first example was that it reduce um + no the first example was that it can regulate temp + the the the general temperature + er it can + even a tree can even make it + make the temp + the the climate for example + a bit more cooler or more humid it depends + and er the second point was that trees um er can + how can I explain this? + um if you + if there are some trees in a city um + the city is less windy + because er trees are able to absorb the um + the wind er + whilst building are not able to do this + er it's with with the buildings + it's it is even worse in fact because the the wind er hit the wall and then go down go around buildings so it's really windy + and um the last point was er that it's quite complicated to + to have trees in city because they use use a lot of a lot of space a lot of room + so it's quite hard to find new places to + for trees + yes and I think that's it.

great + thanks for that + um so what do you think the main topic of this + or topics of this um lecture were?

it's um um + to analyse which are the advantages advantages to have trees in cities and how which could be the solution er to to make this idea possible to + I think.

OK was there any part of the + of the lecture that you found particularly difficult to understand?+ the beginning or the middle or the end?

no + no it was quite OK + it was OK.

you found it quite OK?

yeah, yeah, yeah + it was clear.

OK +um compare this with the last one + the one about the sharks.

yeah.

which did you find more difficult?

the first one + definitely.

why?

because um with this this tests + it it is um + how can I? + it's + I have I have a lot of more stress with this sort of test + because you um you don't want to miss any answer + but with this technique it's it is different + um even if you miss something er + you you you will understand the the general idea what is talking about + but in this test + sort of test um you + when you miss + miss the point then you get you get stress and then for er + for the following questions it's harder + and so it's quite difficult.

do you think you were behaving differently as a listener in the second one?

were you doing things that were different from the first one to the second?

if if what? + I'm sorry.

41 *were you behaving differently as a listener when you did the first test to the way*
42 *you behaved in the second? + were you listening in a different way?*
43 um yes definitely + there's not the same way yeah + because in the first part
44 I'm just focusing on words not the general meaning + but in the second I focus
45 on the the the general ideas the most interesting points.
46 *OK + and did you find one of the recordings more difficult than the other?*
47 the the first one.
48 *you thought the first one?*
49 yeah.
50 *why was it more difficult?*
51 + I don't know really if it was the the speed um definitely again I think this is
52 ++ the way how I + in the second one + it's really easy just to take notes to focus
53 on the main points + but there it's + no definitely I prefer the (unclear)
54 *you say + it's difficult to say if one recording is more difficult than the other +*
55 yeah really.
56 *+ because it was the task +*
57 yeah because of the task.
58 *+ you found*
59 exactly + yeah yeah yeah.