BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

THE EFFECTS OF SINGLE AND DOUBLE PLAY UPON LISTENING TEST OUTCOMES AND COGNITIVE PROCESSING

AR-G/2015/003

John Field, University of Bedfordshire

ARAGs RESEARCH REPORTS ONLINE
SERIES EDITOR: VIVIEN BERRY
ISSN 2057-5203

ABSTRACT

The convention of allowing test-takers to hear a recording twice is a controversial topic within L2 listening assessment. Arguably, it compensates for the lack of visual and contextual information when an audio recording is used; but it can also be argued that, in most real-world circumstances, listeners have only one opportunity to make sense of what is said. Hearing a recording twice has been shown to increase scores, fostering the impression that it renders a test 'easier'. However, little is known about the precise effects of double play upon test-taker behaviour, the focus of this study.

Pre-sessional students (N=73) took a retired IELTS listening test which featured either a multiple-choice format or a gap-filling one. They assumed that they would hear the recording only once, as is customary in IELTS, but were then allowed a second hearing. Scores after the first play were compared with those after the second. There was a general increase, although it varied considerably across individuals. Analysis showed that incorrect or blank answers were replaced by correct ones in relatively few cases. Much more frequent was the persistence of correct answers or incorrect ones. An important issue was whether double play advantaged any particular proficiency level, and thus potentially eroded scoring differentials. In fact, test-takers were found to benefit regardless of level. The data was also analysed to establish whether test format was a factor in the increased scores. Scores from the constructed response task improved more after a second play than those from the selected response one. However, this was mainly because the test-takers experienced difficulty in providing correct answers to gap-filling items after only a single play.

A second phase of the study consisted of face-to-face encounters with individual participants (N=36), who took the same IELTS tests and provided verbal reports on how they had arrived at their answers. This was followed by a semi-structured interview in which participants were asked about their experience of the double play presentation and about how they had made use of the opportunity of hearing the recording twice. The data obtained suggested that, for the majority of the participants, cognitive behaviour differed markedly when they were permitted a second hearing. Important features were a reduction in listening anxiety and greater familiarity with the recorded material, which made it easier to locate the required information. Even on the second play, there was still a heavy attentional focus on word-level decoding in order to check or add answers. But, in the case of many participants, this was accompanied by a wider perspective on the content of the recording and the speaker's goals. These participants managed to acquire the kind of overview that is an important component of academic and professional listening.

Conclusions are drawn as to the validity of employing double play in future listening test design. Specific reference is made to the structure and needs of the Aptis listening test.

Author

John Field is Senior Lecturer in the CRELLA research unit at the University of Bedfordshire, UK. He is especially known for his work on second language listening and his *Listening in the Language Classroom* (CUP, 2008) has become a standard work in the field. His background in psycholinguistics (on which he has written widely) informs much of his thinking and he is currently applying it to the notion of cognitive validity in L2 testing. He undertakes research into how the conventions, input, formats and items used in listening tests shape test-taker behaviour. He also advises test providers on approaches which reflect the nature of the listening skill more accurately and enable it to be assessed more validly. In another life, he was a materials writer and teacher trainer: writing coursebook series for Saudi Arabia and Hong Kong, radio programmes for the BBC World Service and TV programmes for the Open University of China. He continues to advise publishers on materials design.

Acknowledgements

I would like to express my sincere thanks to the British Council for their faith in granting me and the CRELLA Institute the first of the awards in the Assessment Research scheme, and for their patience and support in what has proved to be a long and quite complex project. I would especially like to thank Mina Patel for the personal interest she has shown in the project and the considerate and efficient way she has administered it on behalf of the Council.

The project would have been impossible without the support and hard work of my three research collaborators: Colin Campbell and Jonathan Smith, who collected and collated the group test data, and Sheila Thorn of the Listening Business, who painstakingly transcribed the verbal reports.

I also owe a great debt to the International Study and Language Institute of the University of Reading, where the data collection took place. My sincere thanks to its director, Professor Ros Richards, for her commitment to research within the Institute, and to Dr Sarah Brewer for the considerable help she gave me in overcoming hurdles and ensuring that things went smoothly.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Rationale

IELTS, the most widely taken test of academic and professional language proficiency, allows candidates only one hearing of each recording. In this respect, it differs markedly from (for example) the listening papers of the main Cambridge English suite, where there is a tradition of permitting two plays of the recording. The arguments put forward for the single play specification are that: a) it has always been a hallmark feature of IELTS testing; b) it is ecologically valid in that only one hearing is permitted in a real-life listening event such as an academic lecture; and c) it ensures that the listening test remains a manageable length.

However, little is known about the extent to which the difference between a double play and a single play potentially affects listening behavior, or about the use that candidates make of the double play convention adopted in most international listening tests. On the one hand, the double play might be found to weaken the cognitive validity of a task by eliciting processes unlike those of a real-world listening event; on the other, it might simply serve to compensate for the lack of contextual cues, without markedly changing behaviour.

This has implications for the present format of the Aptis listening test, which (exceptionally amongst tests of its kind) takes advantage of its computer delivered mode to allow the candidate a free choice of whether to opt for a single play of the recording or to hear the recording twice (British Council, 2013). The candidate is not awarded any bonus marks for getting correct answers on the basis of a single listening, so the question arises of whether some candidates may be disadvantaging themselves in opting for only one listening. Conversely, it is possible that candidates may not, by and large, benefit from hearing a recording twice – in which case the facility is redundant, and is adding unnecessarily to the running time of the test.

With the forthcoming revision of the Aptis listening test in mind, it is therefore timely to examine whether the two different listening conditions materially affect candidate behaviour and candidate scores. Concrete evidence for or against either approach might be cited in support of whichever policy is adopted by Aptis, and might contribute usefully to any future rationale for the test's design.

The issue of single versus double play is also likely to become a contentious issue in wider discussion of listening assessment. The difference in this regard between IELTS and other Cambridge English listening papers will increasingly come to the fore as CAE or CPE results become accepted as alternative indicators of ability to perform successfully in an academic context. Some candidates might even choose to sit for other exams in preference to IELTS if they feel that hearing the recording twice will provide additional support and enable them to achieve a higher score.

2. THEORETICAL BACKGROUND

Opinion is divided as to the value of featuring a double play in tests of listening. There are a number of arguments in favour (Field 2008a, p. 159). The first is that playing an audio recording twice compensates for the lack of visual and contextual cues that would be available to a listener in most real-world environments (phone conversations and radio broadcasts being obvious exceptions). Those cues include the speaker's facial expression and gestures, as well as sight of the environment in which an encounter occurs. They also include the movement of the speaker's lips, which have been shown to support phoneme-level perception (McGurk & McDonald 1976). The double-play convention thus counterbalances the fact that an absence of visual input renders the circumstances of a test more demanding than those of most real-world listening events.

The second argument is that the initial play of the recording allows the candidate time to normalise (Pisoni 1997) to the voice, speech rate and accent of the speaker. This process must inevitably be more demanding for second language listeners because of their more limited experience of adjusting to speaker-specific features such as speech rate, voice pitch and accent, while at the same time trying to recognise L2 lexical items. A third, more recently presented, argument relates to the artefactual nature of the task that learners have to perform in a listening test. Field (2013, p. 127) puts forward the view that the operation of checking and matching written information against auditory input is far more complex than most real-world listening and, moreover, that it quite often focuses the listener's attention narrowly on processing at word and clause level. He suggests that a second play permits the listener to achieve this type of task more efficiently and to take account of wider patterns of meaning.

Against the convention, there is a received view among practitioners that it makes the listening task 'easier'. It has indeed been shown (Berne 1995; Cervantes & Gainer 1992) that scores improve when test-takers are allowed to hear a recording twice. But this rather general claim deserves greater exploration. Do scores improve in the case of all individuals? What types of information contribute to an increased score after a second play? Do scores improve across the proficiency continuum or only at certain levels? Above all, are the improved scores more or less representative of the listeners' ability to perform in the real world: given what has been said about the unnatural nature of listening tasks, do these scores increase or reduce the predictive validity of the test?

Currently, the most frequently expressed argument against the double-play convention is that in most real-world listening events, a listener only has one opportunity to hear and make sense of an utterance. The point is not strictly true in all cases: for example, conversational listeners can appeal to their interlocutors for clarification if problems of understanding occur. In addition, the expansion of resources on the internet affords many new possibilities for multiple hearing (even of radio and TV material), and it is becoming increasingly common for tertiary institutions to make lectures available online to students after they have taken place.

A cognitive argument against a second play was put forward by Buck (2001, p. 172) and repeated by Fortune (2004). Buck asserts that a critical component of successful listening is the ability to process information automatically, and that the construct being tested is therefore compromised if a test-taker is allowed to hear a recording twice. The first assertion is certainly true, but these commentators do not appear to fully appreciate that the term *automaticity* relates to the ability to map from word form to word meaning rapidly and with minimal attention load. Instead, they seem to identify the concept narrowly with the ability to identify words in connected speech, which is not quite the same thing. So far as word recognition is concerned, the fact is that much listening, even in a first language, is approximate – as Buck himself recognises (2001, p. 171). Slips of the ear are not uncommon (Bond 1999) and the recognition of a given word or group of words is often retroactive (Grosjean 1985),

occurring up to three or four words later. If there is a case to be made against second play, it does not directly concern automaticity. It is that hearing a recording twice somewhat reduces (though certainly does not eliminate) the extent to which a test measures a learner's ability to form and revise tentative word matches in on-line fashion as a recording proceeds[1]. A second play might conceivably allow opportunities that might not otherwise occur for rethinking initial word matches.

Also cited against double play are certain practical considerations that testers may wish or need to take into account – especially concerning the extra time that double play adds to a test, and the consequent increase in costs. From the perspective of testing professionals, Geranpayeh & Taylor (2008) provide a thoughtful discussion of the issues and conclude that a convincing case can be made for either single or double play 'depending upon factors such as test purpose, cognitive demand, task consistency, sampling and practicality' (p. 2). For another thoughtful review of the topic, see Elliott & Wilson (2013, pp. 197–201).

The issue of single versus double play is especially pertinent when establishing the cognitive validity (Glaser 1991) of a listening test – the extent to which the cognitive processing in which candidates engage shares features in common with the processing that they would employ in a target real-world context. The most obvious hypothesis (on the basis of the 'one-opportunity' argument just mentioned) is that giving candidates two hearings of the listening material distances them from the constraints of a real-life listening event.

However, there is evidence to suggest that the truth is not quite so simple. The effects of double play upon learner and candidate performance have been very little researched but a small-scale study by Buck (1990) concluded that learners engaged with the input in different ways during a second play as compared with a first. Buck's view was that they tended to listen at a local level during a first play but that, during the second, they operated at a more global level. If this is the case, then it would seem, ironically, that it is mainly during the second play that test-takers perform in ways that are closer to the type of behaviour required of them during a real-world situation when they have to follow a lecture, broadcast or an extended conversation.

With Buck's findings in mind, Field (2009) carried out a pilot study in which participants across a range of IELTS scores were asked to undertake two retired IELTS tests featuring different test formats. The findings were of necessity tentative since only 12 individuals were tested. However, the data suggested that test method might be a factor in the extent to which a second listening did or did not advantage an individual. There were also indications that participants with IELTS scores between 5.0 and 6.5 benefited most from the second play, while those in higher and lower ranges did not. The evidence was not sufficient to fully support the latter finding, as the number of participants tested outside the central ranges was quite small. However, it seemed to point towards the hypothesis that low-proficiency listeners succeed in decoding too little of what they hear for a second play to be of benefit to them, while higher level listeners get the answers right the first time and have no need to hear the recording twice.

In a study of the cognitive validity of the lecture listening section in the IELTS test, Field (2012) concluded that the use of a single play appeared to intensify the use of test-wise strategies. The latter often proved counter-productive, with candidates listening for pieces of information that were long past and thus missing subsequent ones.

In a similar study on the processes employed by IELTS candidates, Badger et al. (2012) noted the same phenomenon, but came to a different conclusion, attributing the excessive use of test-wise strategies to the availability of written information contained in the pre-set test items: '…our preferred solution is to maintain the convention that candidates only listen once to the spoken text but without seeing or hearing the questions' (p. 477). This suggestion works well for short recorded clips of up to a minute followed immediately by a single question. But it is very much open to challenge if longer

[1] The revision of initial word matches is certainly an issue for L2 listeners; Field (2008b) demonstrated a reluctance by intermediate-level learners to rethink incorrect lexical segmentations.

recordings are used, as they are in most international tests of listening. There has long been a preference among test designers for pre-setting questions (even at the risk of test-wise strategies) on the arguments that: a) it enables test-takers to listen in a targeted way (which they clearly need to do in a single-play format); and that b) answering a whole set of post-set questions imposes memory demands that are not necessarily part of the listening construct.

While the Badger et al. conclusion is open to challenge on these grounds, it does not obviate their original point – that by pre-setting test items one provides the test-taker with raw material for a whole range of test-wise strategies. The strategies include: mapping from words in the items to similar words in the recording; listening for paraphrase; and forming an advance schema for the content of a recording by using the cues provided by the items. What makes the phenomenon of particular concern in terms of construct validity is that the source of the information is in written, not oral, form.

Sherman (1997) pointed out that, in a double play format, there exists an alternative to the convention of presenting test items at the very outset – namely to present them *between the two plays*. Sherman's study investigated the possible effects upon scores of the point at which test items are made available. In all, four conditions were contrasted: one where written items were provided in advance of the two plays; one where they were provided between the plays; one where they were post-set; and one where there were no conventional questions at all. Interestingly, when item presentation was, as Sherman put it, 'sandwiched', the scores proved to be significantly higher than in other conditions.

A superficial analysis of this result might conclude that the condition weakened the rigour of the test by making it somehow 'easier' than it would normally be. But one surely would expect the circumstances to be *more* demanding, not less, when test-wise strategies cannot be employed during the first hearing (it is worthy of note that the participants themselves expressed a preference for receiving the questions at the outset).

The major conclusion drawn by Sherman was that pre-setting questions does not provide the advantage that is often assumed. An alternative explanation might be that, in the sandwiched condition, the test-takers were listening to the first play in a more natural way – one that was not distorted by prior expectations or by close listening at word level for possible item-recording matches, and that it was for this reason that scores improved. In line with this interpretation, the Field (2012) study provided evidence that, faced with the joint challenges of demanding formats such as gap-filling and a single play requirement, test-takers largely processed spoken input at low perceptual levels. This type of processing was, Field argued, not sufficiently representative of real-world academic listening.

Field (2013, p. 127) went on to link test-taker behaviour more specifically with the types of format favoured in international tests of listening:

> Less often cited in favour of double play is the nature of the listening task that the test-taker is asked to perform. Conventional test formats…require the candidate to undertake a process of checking or matching information. This type of activity is an artefact of the testing situation rather than a characteristic of most real-world listening; it is therefore surely fair to allow the test-taker the means of accomplishing the task as efficiently as possible. The limited research that has considered the effects of double play (Buck 1990; Field 2009) suggests that different types of checking process are involved in the two hearings. The test-taker makes use of the first play to establish the location of information in the recording and to make a preliminary match against the items in the paper. Processing is thus very much at the level of *lexical search* and *parsing*. On a second hearing, the test-taker is able to locate the relevant information more confidently and to confirm or revise provisional answers. In addition, he/she moves on to construct a higher-level *discourse representation*. While this may not mirror real-world listening, the processes involved seem entirely appropriate to the artificial nature of the task that has been set.

Buck (2001, p. 171) similarly draws attention to the artificial nature of a listening test:

> The testing situation is unnatural in demanding that the listener comprehend with a much greater degree of precision than is normal. In other words, in preserving the situational authenticity by giving the text only once, we are sacrificing interactional authenticity by asking listeners to understand more precisely than in the target-language use situation. Given this, playing the text a second time does not appear such an unnatural thing to do.[2]

Mention of task conditions raises the interesting possibility that the differing demands of test formats may be a factor in determining how much benefit a test-taker derives from hearing a recording twice. In trials which formed part of a revision project for Cambridge CPE, Boroughs (2003, p. 333) reported that the difficulty of constructed response items increased in a single play condition to a much greater extent than did that of selected response items.

The present project explored these various issues relating to double play in what would appear to be the first full-scale research study of the convention. It replicated the Field (2009b) pilot with a larger and differently distributed group of participants and with the needs of the Aptis test in mind. One particular area of interest was the impact of double play upon the scoring system and whether it rendered the system more or less transparent. Another was the possible effect of test format upon the use made of the second play. Here the comparison was between a gap-filling format and one employing multiple-choice question (MCQ) items.

A broader consideration was to investigate whether the double play convention conduces to a type of listening which differs substantially from that employed in real-life contexts, thus raising questions of cognitive validity. The question was of relevance when determining the conventions to be adopted in the revised Aptis listening test. If the revised test followed the usual procedure of a double play, it would be useful to be able to justify the choice on the basis of empirical evidence (something that other tests have never done) and to defend it against the suggestion that different cognitive processes are employed under this condition. Similarly, if Aptis adopted the single play convention, it would be desirable to demonstrate that the decision did not discriminate unfairly against borderline candidates.

[2] Buck's stance is ambivalent. As already noted, he elsewhere (p. 172) argues for a single play on the grounds that what he terms automatic processing is compromised if a test-taker hears a recording twice.

3. RESEARCH DESIGN

3.1 Research questions

RQ1: To what extent does a candidate's score improve with the opportunity to hear a recording twice?

RQ2: Is level of proficiency a factor in determining whether candidates benefit from a second play in terms of an increased score?

RQ3: Is test method a factor in the way in which a second play is handled?

RQ4: In what ways does a candidate take advantage of the second play to: a) check; b) change; and c) supplement answers?

RQ5: Does a candidate behave differently in cognitive terms when it is established that a second play of the recording will be available?

3.2 Methods

Two methods were employed. The first entailed testing participants in class and volunteer groups to provide quantitative data from a larger population than could be tested individually. It provided data in the form of test scores and answer sheets that could be analysed. The second method tested participants singly to enable them to provide verbal reports (Ericsson & Simon 1993) of the processes they had employed in arriving at their answers. This produced quantitative data in the form of scores and completed answer sheets, as well as qualitative data in the form of protocols.

Group testing

In order to strengthen the statistical data for answering RQs 1 and 2, participants were tested in volunteer and class groups. They were asked to listen to, and answer, a lecture-based section from a retired IELTS test, featuring one of two test methods (either MCQ or gap-filling). Three groups (N=40) were tested on the section featuring an MCQ format; and two groups (N=33) were tested on the section featuring a gap-filling format. These participants were not told in advance that they would be permitted a double hearing of the recording; they were allowed to assume that the task would be on a single play basis, as is usual in IELTS. They were provided with red pens to use during the second play so that their answers could be distinguished from those given during the first.

One-to-one testing

In addition, a separate set of individual participants (N = 40) was asked to listen to, and answer, questions from one of the same two retired IELTS tests. As a preliminary step, they were asked to undertake a listening task from a retired IELTS paper, where only one play was permitted. This was to provide them with a point of comparison when they came to undertake the double play task. Before the double play task, they were told that they would hear the recording twice.

Pauses were inserted into the recordings, to enable participants to report their answers so far. During the first play, they were asked their reasons for choosing a particular answer. During the second, they were asked if they had made any changes to their original answers or had added to them, and why they had done so. Like the group participants, they were asked to use red pens to keep second-play answers distinctive on the answer sheets. Finally, they were briefly interviewed about how they had behaved during the second play as compared with the first.

Self-evidently, verbal report in listening has to be retrospective. It consequently runs the risk of memory effects. The problem was obviated here by dividing the recordings into sub-sections and asking respondents to report and justify only three or four answers at a time. The tapescripts in Appendix 2 indicate where the pauses fell. The recency of participants' decisions ensured more accurate reporting of the processes entailed in arriving at them, while the form of words in the chosen answers served as a memory cue and thus produced stimulated recall along the lines envisaged by Gass & Mackey (2000).

3.3 Participants

All participants were drawn from pre-sessional intakes at the International Study and Language Institute (ISLI) of the University of Reading. Those participants tested in groups (N=73) were controlled to ensure they had no previous residence in an English-speaking country and were tested within six weeks of arrival. Because they were volunteers or in their usual class sets, it was not possible to control in advance for their pre-entry IELTS listening scores. The participants' scores ranged from 4.0 to 7.5, but were more heavily weighted towards the lower end.

The L1s within the group tested on the MCQ materials covered a wide range with Arabic predominating: Arabic 15, Mandarin Chinese 8, Kazakh 4, Thai 4. Hindi 2, Japanese 2, Polish 2 and 1 each for Farsi, Bahasa Malaysia and Russian. Within the group tested on the gap-filling materials, Mandarin Chinese predominated: Chinese 15, Arabic 6, Kazakh 8, Spanish 2, and 1 each of Japanese and Vietnamese.

The individual participants were similarly controlled to ensure that they were naive listeners in the sense that they had had no previous residence in an English-speaking country and they were tested within five weeks of their arrival in the UK. They were also controlled for levels of listening proficiency to ensure that their IELTS listening scores prior to being accepted on the pre-sessional course covered a complete range from 4.5 to 7.5 with a higher representation in the central bands. These participants mainly had Mandarin Chinese as their L1 but three had Thai, two had Arabic and two had Kazakh.

3.4 Materials

Two sections were chosen from retired IELTS listening papers. The reasons for choosing IELTS material were firstly that it has been reportedly designed for the single play condition and secondly that the target population were familiar with the IELTS formats. In addition, IELTS is designed to measure proficiency across a wide range of levels rather than a single one; and aims, like Aptis, to measure potential performance in professional and academic contexts.

The two sections used were taken from the lecture-based Section 4 of the IELTS test (*Cambridge IELTS 6*, p. 131 and p. 137). One section represented the MCQ format and the other the gap-filling one. They were chosen partly because of the relative concreteness of the keys targeted. A major consideration was that both were narrative texts and, therefore, not too informationally complex; and also that both featured lecture-style delivery at a speech rate which was middle range for that type of discourse. One section (henceforth Text A) concerned the history of early moving pictures and was delivered in a General American accent. The other (henceforth Text B) concerned the history of the East End of London and was delivered in a Standard British accent. The answer sheets and tapescripts for the two sections chosen appear in Appendix 1 of this report.

One difficulty in employing this material for the present project was that IELTS item writers tend to favour a mixed approach to test format, even within a section. So, while seven items of Text A employed an MCQ format, three featured gap-filling. While seven items of Text B employed a gap-filling format, three featured multiple-matching. When discussing the impact of format upon test-taker performance (RQ3), the figures quoted will therefore refer to only seven out of the ten items in each section.

3.5 Procedure

Participants in groups were allowed to assume that they would be undertaking the test under normal IELTS conditions (i.e. with only one hearing of the recording). After the first play, they were then told that they would be able to hear the recording again. The aim was to monitor the scores achieved under what were assumed to be single play conditions against the final scores achieved after the additional play had been unexpectedly provided.

Individual participants were told before the task that they would hear the recording twice, so that their behaviour might be studied in circumstances where they knew that a double play was allowed. These participants were first tested on another retired IELTS test under single play conditions. The original aim was to see to what extent they tackled the task differently in the knowledge that they could only hear the recording once. In the event, it proved hard to conclusively identify marked differences of behaviour during a single play as compared with the first hearing of a double play. This line of enquiry was therefore not pursued – but the single play nevertheless provided a useful benchmark against which the participant could be asked to compare the double play experience.

Ethical permission was sought from the ISLI, University of Reading, where participants were studying. Individual participants were asked to sign a consent form that complied with the university's ethical guidelines; a copy appears in Appendix 2 of this report. Each was paid £10 for their time. Group participants were asked to provide formal ethical consent at the head of their answer sheets (see Appendix 1).

There were certain complications affecting the collection of data at the University of Reading in what was a transitional period during which the International Study and Language Centre became a full Institute of the university, with consequent uncertainties about regulations and procedures. There were also problems caused by building work, which complicated the pre-booking of interview rooms. These unpropitious circumstances considerably delayed some of the data collection.

4. QUANTATIVE DATA ANALYSIS

Scores were calculated for each participant in both the group and the one-to-one conditions. They were inserted into spreadsheets which contained information about a candidate's IELTS score for listening before arriving in the UK, the candidate's score on the first play, the score on the second play and the differential between the two. All information was coded to ensure participant anonymity.

The scores logged were out of 10 as each section of the IELTS listening test consists of 10 items. However, as already noted, the sections tend to use mixed formats. A further set of scores was therefore calculated, representing performance on a single format (MCQ in the case of Text A and gap-filling in the case of Text B). These scores (out of 7) provided evidence of the possible effects of format on performance.

4.1 General impact of a second play

The analysis will focus on the group participants (N=73), who undertook the listening tasks under test conditions. This subset (nearly two-thirds of the population tested) had not been told in advance that they would be able to hear the passage twice – thus eliminating the likelihood that some of the score increases might be due to their delaying a choice of answers until the second play.

Scores following Play 1 and following Play 2 were totalled for participants. Table 1 below shows the results, together with mean scores and SDs.

Table 1: Scores for group participants (N=73)

	Responses Play 1	Responses Play 2	Differential
Total correct	272	347	75
% correct	*37.26%*	*47.53%*	*10.27%*
Mean score (SD)	3.73 (2.06)	4.75 (2.01)	1.03 (1.37)

It will be immediately obvious that a second play of the recording led to an improvement in overall scores. A z-test for proportions showed the difference between Play 1 and Play 2 to be highly significant: $z = 3.97$, $p<0.01$. At the same time, it is important to note that:

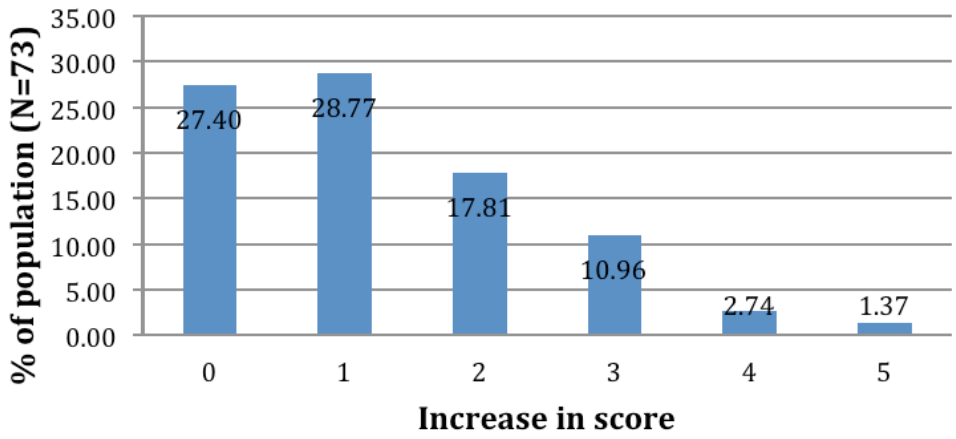
- even after Play 2, less than 50% of responses were correct
- the score increase as a proportion of Play 1 scores was only 27.56%.

It would seem clear that the opportunity of hearing a second play has the general effect of increasing comprehension but that it does not do so by the margin that might be assumed.

A degree of variation between the participants was indicated by the relatively high SDs in Table 1. Closer inspection indicated that, over the 73 participants, 27.40% showed no change in score as a result of the second play while 61.65% increased their score to different degrees. Only a minority (8 out of 73) found themselves with a worse score after the second play. Of these, seven lost only one mark and one lost two. It would seem abundantly clear that the opportunity of hearing a second play has the general effect of increasing comprehension rather than perpetuating incomprehension.

Figure 1 below shows that most individuals tested in groups increased their scores during the second play and did so by anything between 1 and 5 points. However, what is perhaps most striking is the proportion of participants (56.17%) who showed no increase in score or an increase of only 1.

Figure 1: Increase in scores of group participants



4.2 Decision-making during the second play

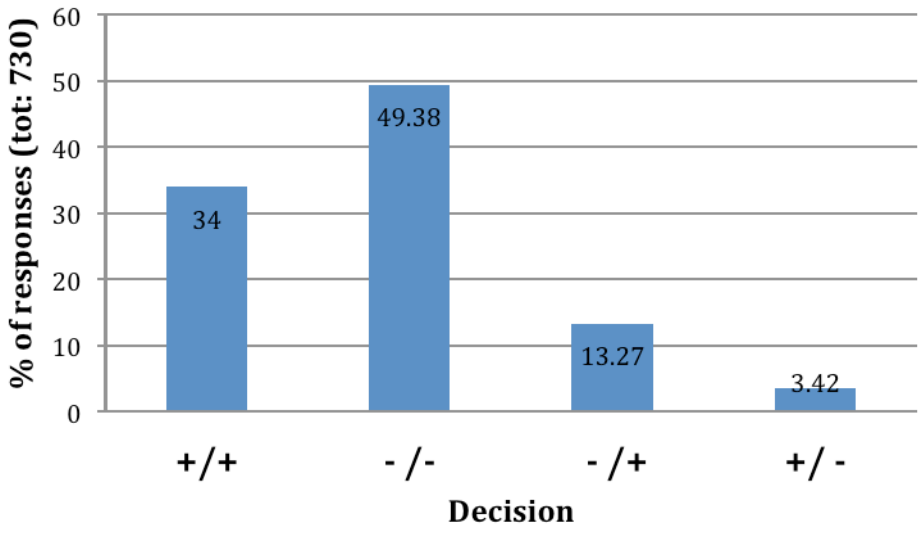
It would seem, then, that hearing a recording for the second time enables most test-takers to replace blanks with correct answers, or to revise incorrect answers. However, the possibility must also be addressed that an overall increase in scores conceals the fact that some candidates, far from benefiting from the double play, took a backwards step and replaced a correct answer with an incorrect one. An analysis was therefore made of the changes that had occurred during the second play: these related to the number of items (correct and incorrect) that were left unchanged, the number that changed from correct to incorrect and the number that changed from incorrect to correct (or indeed from blank answer to correct).

Answers given by group participants during the first play were cross-compared with the answers given during the second. The decisions made were classified as follows:

- **+/+** correct answer left unaltered
- **- / -** incorrect answer left unaltered; blank left unfilled
- **- /+** wrong or blank answer changed to correct one
- **+/ -** correct answer changed to a wrong one.

They are shown here as percentages of the total responses.

Figure 2: Decisions made by group participants (N=73) during second play



It would seem that in many cases the second play provided test-takers with a means of checking a response that was already correct. An even more common occurrence, however, was that a participant left an incorrect answer or a blank slot unchanged despite having the opportunity to check the recording again. One can imagine a number of scenarios. One is that the recording was not understood during the first play and that a second hearing did not serve to clarify things at all. Another is that some of these participants left a blank during the first play which they filled with a random answer during the second. A third (with implications for cognitive validity) is that a guess made during the first play might have gone on to shape a listener's understanding of what they heard during the second play, to the point where they were not capable of judging that it was incorrect.

An unexpected finding was the relatively small proportion of answers (13.3%) where participants took advantage of the second play to change incorrect answers or to identify correct ones for the first time. They were counter-balanced by a smaller number (3.4%) where participants substituted a wrong choice for a correct one. This appears to give the lie to the notion that a second play results in a test becoming considerably easier and even arguably too easy.

4.3 The effects of listening proficiency on performance in double play

So far, the results have been examined in terms of broad patterns of behaviour. A major question that now needs to be addressed is whether listening proficiency determines the extent to which a learner is capable of taking advantage of the double play convention. Results from the pilot reported earlier (Field 2009) suggested that it is mainly candidates with IELTS scores in the central bands (5.5 to 6.5) who benefit from the chance of revising or amplifying their answers during a second hearing.
An obvious conclusion seemed to be that lower-proficiency listeners have such difficulty in decoding the input that they do not succeed in matching enough of what they hear to words, even on a second play. A similar conclusion in respect of higher-proficiency listeners was that they get their answers right the first time and simply use the second play to check them.

The numbers of lower-level and higher-level participants in the pilot study were very small; and therefore these interpretations remained simply hypotheses. However, they raise serious issues of scoring validity. If the use of a second play benefits only one proficiency band of participants, does it unfairly inflate their scores in relation to those of other candidates? On the other hand, (given the disadvantages that a listener suffers from when asked to perform complex tasks involving audio material with no visual cues), could it be claimed that middle-range candidates are penalised unfairly by only being allowed one play when they can perform much better if allowed two? It is, after all, these candidates who are the most vulnerable to fine gradations in scoring because their scores are borderline in relation to the targets generally used for university admission at undergraduate level.

Participants in the present study were therefore grouped into three proficiency bands, according to their pre-entry IELTS record for listening: Low (IELTS Listening 4.0 to 5.0), Central (IELTS Listening 5.5 to 6.0) and High (IELTS Listening 6.5 and higher). The bands were chosen in recognition of the fact (see Appendix 3) that an overall IELTS score of 6.0 is the lowest level accepted by many UK and overseas institutions for university admission. Table 2 shows the number of participants in each band.

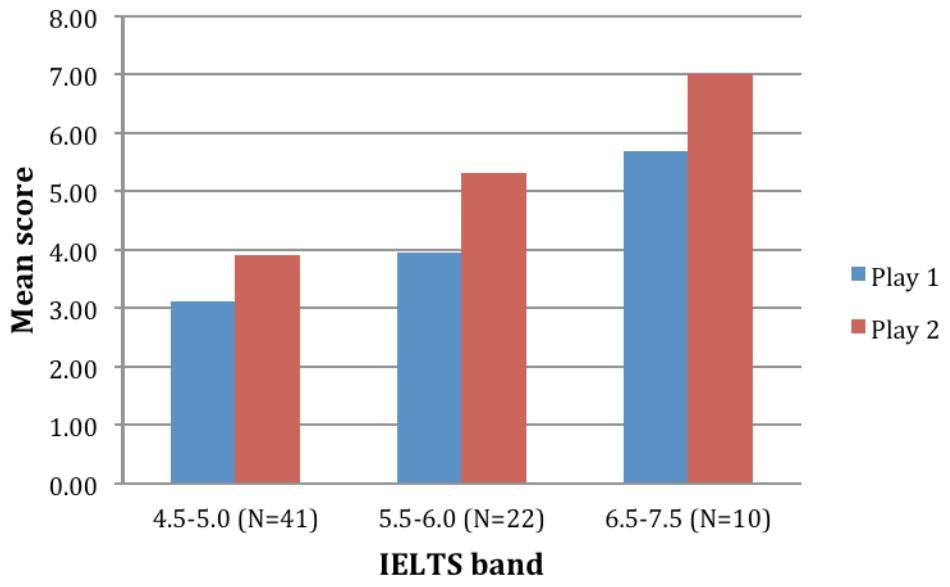
Table 2: Listening proficiency of participants by bands

	Individual (N = 40)	Group (N = 73)	Total (N=113)
Low (4.0 to 5.0)	6	41	47
Borderline (5.5 to 6.0)	20	22	42
High (6.5 to 7.5)	14	10	24

Mean scores within the three bands were calculated for group participants (N=73). Figure 2 contrasts the mean score after Play 1 with that after Play 2.

A z-test for proportions found the differences between total scores on Play 1 and Play 2 to be significant at each of the three levels: for the low band $z = 2.34$, $p<0.05$; for the central band, $z = 2.87$, $p< 0.01$; for the high band, $z = 1.98$, $p<0.05$.

Figure 3: Score change on second play: group participants (N=73)



These findings discredit the hypothesis put forward by Field (2009) that it is chiefly candidates in the central score bands who benefit from the second play. It would appear that candidates in all three bands benefit – and do so in a way that, far from reducing or distorting score differentials, sustain them. Table 3 below demonstrates this point: the penultimate column shows the increase in mean scores for each group following a second play, while the last column expresses the increase as a percentage of the original mean score. Figures in brackets indicate SDs.

Table 3: Proportionate increase in mean scores across three levels: Group participants (N=73)

IELTS band	Play 1	Play 2	change	% increase
4.0 to 5.0 (N=41)	3.12 (1.79)	3.90 (1.73)	+ 0.78 (1.37)	25.00
5.5 to 6.0 (N=22)	3.95 (2.03)	5.32 (1.67)	+ 1.36 (1.40)	34.48
6.5 to 7.5 (N=10)	5.70 (2.00)	7.00 (1.70)	+ 1.30 (1.25)	22.81

It will not have escaped notice that the mean scores achieved by participants on the first hearing of the recording fell well below their attested IELTS band scores, and only rose to become representative of those scores following a second hearing. The recordings and tasks employed in this study were taken from the one section of the Academic IELTS test that focuses on lecture-style material. The remaining three sections feature material of a more interactional kind – namely instructions and advice on matters relating to campus life and/or tutorial exchanges. The four sections are scored equally and it is somewhat of an anomaly that the principal listening challenge faced by intending students (namely listening to lectures) is so little represented in the test. Lecture listening is considerably more demanding than listening to more general conversational or explanatory material. It is likely to entail language that is more formal, with lexis of lower frequency and a greater degree of syntactic subordination. But perhaps more important is the fact that lectures are generally expository or discursive in content. As a result, they tend to be more informationally dense than conversation and thus more cognitively demanding to follow. They are also more conceptually complex, with the listener needing to relate what is said to wider goals and to trace links between the points that are made. In cognitive terms, this requires quite extensive reliance on higher-level processes – placing information in a wider context, drawing inferences, recognising speaker goals, constructing a line of argument and so on.

The evidence cited above suggests strongly the following.

a. Far from blurring the score differentials between candidates (for example, by unduly favouring weaker candidates), the effect of a double play is to increase those differentials.

b. Candidates do poorly on the lecture-listening task under single play conditions. This suggests that many of those who achieve high overall IELTS listening scores do so on the basis of successful performance in the less demanding conversational sections. Such scores can only be weakly predictive of a candidate's ability to perform competently in academic contexts when confronted with lecture-style material.

c. It is only when candidates are allowed the benefit of a second hearing of lecture material that they appear to score at a level which is consistent with their performance elsewhere. The double play enables them to obtain an overview of the information covered and where it occurs – something that arguably would not be available in real life. On the other hand, it also compensates for the absence of what a real-world lecture situation *would* supply in order to compensate for informational density and complexity. This includes the signposts provided by PowerPoint slides and by lecturer gestures and facial expressions, as well as a degree of prior topic knowledge.

There are clearly lessons to be drawn in relation to the revision programme for the Aptis listening test. When Aptis is employed to assess academic listening, input material should be such as to elicit the types of cognitive response that one would expect in real-world performance. This should entail quite a high proportion of lecture-style material, which is considerably more demanding than conversational exchanges of information. More relevantly to the present research focus – double play scores with this type of material seem to be a better indicator of proficiency level than those obtained on the basis of a single play.

4.4 The effects of question format on double play scores

The analysis now addressed the question of whether the benefits derived from the opportunity of hearing a recording twice might vary according to the test format employed. It will be recalled that, of the two IELTS past papers used in the study, one featured 7 multiple-choice items out of 10 while the other featured 7 gap-filling items. The scores for the four groups tested were re-calculated to include only these items; means were derived; and the data was subdivided according to whether a given group had been presented with Text A (MCQ) or Text B (gap-filling).

The resulting mean scores out of 7 are shown in Table 4 with figures in brackets indicating SDs. The figure in the last column is again derived by showing the mean overall increase as a percentage of the first play mean. Figures in italics represent scores rounded up out of 10 to enable comparison with other results.

Table 4: Mean scores out of 7 by format: group participants (N=73)

Format	Play 1	Play 2	change	% increase
MCQ (N=40)	3.22 (1.61) *4.6*	3.75 (1.33) *5.36*	0.52 (1.28) *0.76*	16.28
Gap-filling (N=33)	1.64 (1.98) *2.34*	2.58 (1.97) *3.69*	0.91 (1.13) *1.35*	57.41

Italics: scores recalculated out of 10 for purposes of comparison

There was a striking difference between the two conditions, both in score level after Play 1 and in the mean increase in scores. This could simply have reflected differences between the groups in the range of proficiency levels represented within them. But, in point of fact, the MCQ groups were the *weaker* of the two in terms of proficiency, with a higher number of participants at IELTS level 5.0 and below (12 as compared to 6 in the gap-filling condition) and a lower number at IELTS level 6.5 and above (3 as compared to 7). The result might possibly be attributable to differences in the difficulty of the items or recordings. Alternatively, it seems more likely to reflect the greater cognitive demands of the gap-filling format – a divided attention task which requires candidates to read and write as well as listen (Field 2013, p. 131).

The advantage gained in relation to a second play was markedly greater in the gap-filling condition. Indeed, a z-test for proportions comparing the first and second play results for MCQ failed to reach significance ($z = 1.77$, n.s.) while the same statistic for the gap-filling results was highly significant ($z = 3.14$, $p<0.01$). This difference cannot be attributed to gap-filling participants deciding not to write an answer during the first play in the knowledge that they would have be able to listen again; group participants did not know in advance that they would hear the recording twice.

The results were checked against those for individual participants (Table 5). Here the scores were higher overall. Both increases as a result of the second play were significant (MCQ: $z = 3.44$, $p<0.01$; gap filling: $z = 3.25$, $p<0.01$). However, the gap-filling results again fell well below those for MCQ, supporting the hypothesis of item difficulty and/or a format effect.

Table 5: Mean scores out of 7 by test format: individual participants (N=40)

Format	Play 1	Play 2	change	% increase
MCQ (N=24)	3.96 (1.37) *5.66*	5.21 (1.35) *7.44*	1.25 (1.07) *1.78*	31.58
Gap filling (N=16)	2.13 (0.96) *3.04*	3.63 (1.26) *5.19*	1.50 (0.89) *2.15*	70.59

Italics: scores recalculated out of 10 for purposes of comparison

A t-test of independent means compared the differentials between the two plays in the group MCQ condition with those in the gap-filling condition. The result was not significant: $t(71) = 1.43$. n.s.
It was apparent that this was largely due to a high level of between-participant variation. The numbers of group participants achieving each score from 0 to 7 were therefore compared after the first play and after the second in order to gain a more detailed picture of the impact of double play. The results appear below in Figure 4 (MCQ) and Figure 5 (gap-filling).

Figure 4: MCQ: Group participants (N=40) achieving correct answers

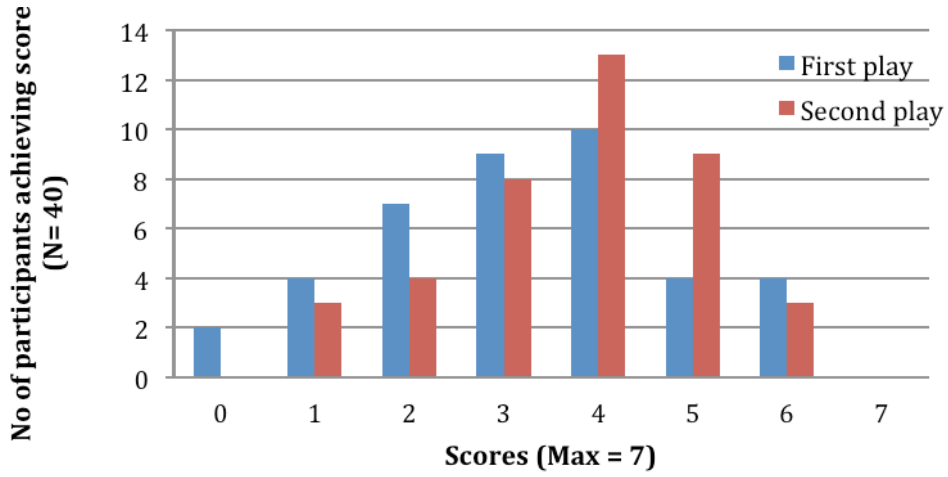
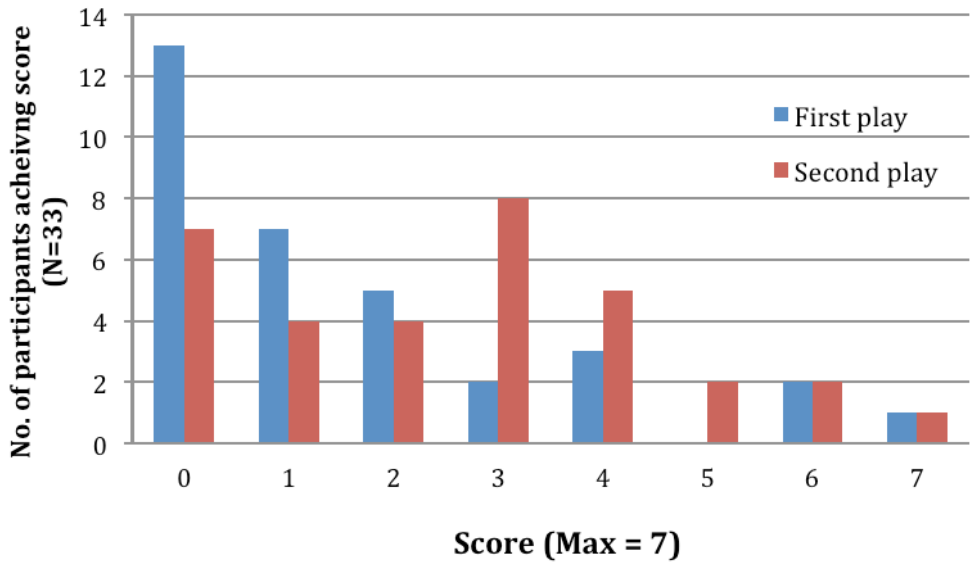


Figure 5: Gap-filling: Group participants (N=33) achieving correct answers



These graphics provide interesting evidence of the way in which test-takers recuperate from an initial low score when they are permitted a second play. But they also suggest that the benefits of a second play may vary according to the format employed – with the greater attentional and accuracy demands of gap-filling leading to greater uncertainty when giving answers or an inability to retrieve sufficient information, and hence to a greater dependence on hearing the recording again.

5. QUALITATIVE ANALYSIS

In all, 40 participants contributed verbal reports; but the general procedure and the questions posed by the researcher were modified after trials with the first four. This section of the report therefore draws upon protocols from 36 of the 40. As already noted, the population sample was based upon a range of IELTS scores, with higher representation in the central bands. The distribution amongst the 36 participants featured was as follows: 4.5 (4); 5.0 (2); 5.5 (10); 6.0 (8); 6.5 (6); 7.0 (3); 7.5 (3). Of these participants, 23 experienced handling Text A (MCQ) while 13 reported on Text B (gap-filling).

The information provided by the protocols was of two types:

- **verbal reports of the processes** in which participants engaged during the first and second plays of the recording. The evidence examined here chiefly relates to the process of checking answers during the second play and the reasons for any changes that were made. In particular, interest attached to the level of processing engaged by the listener: i.e. whether attention was directed at a perceptual level (either involving words or clauses) or at a conceptual one (involving points of information and general line of argument)
- responses **to questions posed by the researcher** about the experience of the second play. The precise wording of the questions was adjusted somewhat to fit the participant's level of English and powers of self-expression. However, in all cases, the questions avoided precise terms such as *anxious* or *focus your attention*, consistently using vague alternatives in order to avoid leading the participants.

5.1 Verbal reports

Following the second play of the recording, participants were asked to report on whether they had changed any of their answers and to give the reasons for doing so. Where their answers remained the same, they were asked to report on whether and how they had checked that these original answers were correct.

It was assumed that participants might follow one of two general courses. They might listen generally on a first play to gain an idea of what the text covered and where information was located; then go on during a second play to focus on details that might supply answers to questions. Alternatively, they might locate provisional answers to questions during the first play at quite a local level, and then go on to check those answers or place them in a wider context during the second play.

Evidence from the verbal reports provided by participants suggested that most, if not all of them, adopted the latter course. It was clear, particularly among participants at higher levels of proficiency, that some of the processing that took place during the second play was less localised than that during the first play. In checking answers, some participants expanded their recognition of what was said to include the immediate co-text of the word or phrase which provided an answer to the question; others reported that they extended their understanding of the speaker's intentions while at the same time focusing locally on parts of the recording which were critical to establishing that their answers were correct.

This is entirely consistent with what is known about the complexity of the cognitive processes entailed in listening to speech. It is well attested that listeners are capable of processing perceptual and semantic information at various levels simultaneously (McQueen 2007). These levels correspond to the components that feature in current cognitive models of listening performance (Field 2013); they include lower-level processes relating to word recognition and clause parsing; and higher-level processes that entail assembling information, extracting implicit and inferential meaning and establishing the overall line of argument. Within this multi-stranded operation, a listener is capable of choosing to focus particular attention upon a single one of these various sources of information.

What is very obvious from the protocols obtained during the study is the extent to which participants targeted *individual words* in the text, even if they were also expanding their understanding of the text as a whole. This was to be expected in the case of low proficiency listeners, who tend to focus heavily on word-level information because of difficulties in decoding speech, but in fact the phenomenon was observed across all proficiency levels except the highest.

The result is perhaps unsurprising in test conditions, where much of the perceptual information that provides correct answers lies at the levels of the word and the lexical chunk. Indeed, there was a strong effect of test method upon the techniques adopted. Participants handling Text A frequently reported during the second play that they continued to listen out for key words taken from the written MCQ options or for paraphrases of those words. Some also reported listening out for single words which would eliminate certain options that they had not fully focused on during the first play. Those who were presented with Text B (the gap-filling task) reported using as anchor points the words identified during the first play. They listened out for them, not only to check that they were correct, but also to add missing words to their responses. Evidence of this format-dependent process was also found in some of the completed answer sheets, where the filled gaps consisted of a word or words inserted during the first play (in black ink) accompanied by an additional word inserted during the second play (in red ink).

Focusing in this way on decoding previously unnoticed or unrecognised words enabled learners to perform several test-specific functions:

- to check words already contained in answers – whether words they had written when filling gaps, or words occurring in a chosen MCQ option or paraphrasing it
- to listen out again for key words that were present in MCQ items or that paraphrased them
- to supply missed words in a partially completed gap-filling item
- to identify words that might eliminate particular MCQ options.

The protocols provided evidence of all these checking process being used, across proficiency levels. This would appear to qualify the assumptions made by Buck (1990) and more recently by Field (2013) that a first play is likely to be more narrowly focused on perceptual processes, while a second one might take greater account of global meaning. The fact is that, under the circumstances of the text and with a less than perfect knowledge of the L2, there are strong constraints upon learners to check their answers closely by drawing upon perceptual cues at word level.

However, it was also noticeable that for quite a number of the participants the focus at word-level served other purposes which stretched beyond simple lexical recognition. It enabled them to:

a. *add words to utterances* that were incompletely decoded during the first play, and thus extend understanding of the recording
b. re-visit *syntactic indicators within an utterance* to check whether they match or eliminate a chosen MCQ option
c. identify words that serve as *locators for particular pieces of information* that are being sought
d. identify words from which *wider information might be inferred*.

The set of extracts below provide examples of the way a single, quite proficient participant (IELTS level 6.5) used newly identified words in order to extend meanings derived during the first play. The recording featured was Text A (see Appendix 1).

(1) R: *what did you do? Did you check it?*
S: *yes because I heard um+the the lecturer emphasise on the word <u>all</u>* [Oa:109-10][3]

(2) R: *is there anything you heard this time + which made you sure that it was point B?*
S: *er yes also the same with the first time + er the fellow is quite clever and he can use some + some technology to make… or create some new system* [Oa: 120-122]

(3) R: *you changed to the point C… why?*
S: *yes because um I I know the reason why I picked the wrong answer A + because there is a word film but A is a passive not active + so not be filmed + it is er + what's said is the camera is too heavy + about how much + how many kilograms*
R: *how many kilograms? + did you know?*
S: *+ er thousand? + I don't know for sure* [Oa: 129-138]

[3] Participants were coded to ensure anonymity. Individuals are identified by one or two upper-case letters (here O). This is followed by a lower-case letter indicating the text they were presented with (a refers to Text A). Numbers refer to turns in the recording.

(4) R: *what about number four?*
S: *er yes the point A is the correct answer because er + the the woman said heard + the word heard…+ so heard is that have the similar meaning of being told about* [Oa: 142-144]

(5) R: *you said you heard the word tension + and that made you decide that A was right*
S: *yes the meaning is there are two reels… and the two reels are so + so connected with each other + and the the woman said er + there is a tight + there is a tension between the two wheels …*
R: *and what happened because of the tension?*
S: *I don't know*
R: *OK + but never mind + you got the words tension + and you thought it was something to do with taking away… the tension by adding this third reel*
S: *yes* [Oa: 187-200]

- Extract (1). Text: *all four feet off the ground*. A typical example of word-level checking, with the participant listening for a perceptually prominent word that confirmed her choice of option.
- Extract (2). Text: *asked a young Scotsmen in his employ to design a system, which he did. Now this young fellow was clever because the first thing he did was study other systems*. The participant pieces together a proposition based on three prominent words (*clever, technology, system*), which confirms her choice of option.
- Extract (3). Text: *The camera weighed over 200 kg and only one person at a time could see the film*. The participant focuses attention on the syntactic structure *could see the film* and is able to reject Option A: *Only one person could be filmed*. She then focuses on the other problem mentioned, and, although she could not report the exact weight of the camera, she inferred from the words *weighed* and *kilograms* that it was heavy.
- Extract (4). Text: …*rival European systems started to appear once people had heard about it*. The participant matches the sequence *had heard about* to the MCQ paraphrase *had been told about*.
- Extract (5). Text: *the Lantham Loop…was the simple addition of a third reel between the two main reels, and this took all the tension away with the result that the film stopped snapping*. The participant identifies references to *third reel, two reels* and *tension*. On the basis of this, she is able to identify the correct Option A: *Removing tension between the film reels*, even though she is not fully aware of what occurred.

It can thus be said that, although superficially this type of processing was heavily targeted on word recognition, it also served the wider purpose of expanding the body of information that the listener had built up during the first play. The second play clearly provided additional opportunities for strategic listeners but it is entirely fair to reward the ability to extract accurate information from a partially understood text, which is an important asset for an academic listener.

The point should also be made that, alongside the word-level processing, 10 of the 36 participants showed some signs of operating at levels above the word. Higher proficiency individuals, in particular, were capable of producing chunks of text or of reporting complete units of information in support of the answers that they had chosen.

Here are some examples from one participant (Qb):

(6) I did listen to see if I heard the feed the people again (Qb:71)

(7) I double-checked + I I still think I heard the metal and leather goods (Qb:81)

(8) because I heard that there were not enough room for people…to live (Qb:115)

This particular respondent had an IELTS listening score of 7.5 but, at lower levels, there was also occasional evidence of focus on chunks. One IELTS 5.5 participant was able to quote text that was nearly verbatim:

(9) um + the crops and the livestock um help feed the people (Hb: 91)

5.2 Interview questions

Four questions were addressed to individual participants. They were on the following lines, though the precise wording was varied somewhat in response to the participant's level of English and powers of self-expression.

1. Did you feel more comfortable when you were able to listen to the recording twice?
Immediately before the double play task, participants had been exposed to the experience of answering questions on another recording, heard once only. This provided them with a yardstick against which to make a comparison. The vague word *comfortable* was used throughout with a view to eliciting a range of unbiased responses, which might relate to listening anxiety, perception of recording difficulty, familiarity with speaker's voice and accent etc.

2. Did you feel that you listened differently when you heard the recording the second time? Or was it similar?
Again, the vague word *differently* was used in order not to bias responses, which might relate to familiarity with speaker's voice, greater familiarity with text, perception of speech rate and recording difficulty, level of attention and so on. Especially of interest was the hypothesis that the first recording might be used by participants in a very focused way for identifying answers, while the second permitted them to gain a wider understanding of the content, including an awareness of argument structure and the relative importance of different points

3. Did you find it easier to understand the recording when you heard it the second time? Or did you just check your answers?
It was assumed that the general answer to this question would be 'yes' but that some participants with higher IELTS listening scores might have understood the recording well enough on the first play not to need a second. Here, again, there was the possibility that participants might comment on efforts of attention or on their changed perceptions of the difficulty of the recording, text or items

4. What did you pay attention to when you listened for a second time? Words or ideas?
This aimed to elicit participants' own impressions of their listening behaviour, which could be matched against the evidence provided by the verbal reports. It was followed up where necessary with probes as to the type of unit (word, point of information, line of argument) that the participants believed they had principally focused on.

The responses to questions given by participants were examined and coded into a number of themes: the effect of familiarity, listening anxiety, level of attention and information extracted during first and second plays.

5.2.1 Familiarity

Of the 36 participants, 24 commented in one way or another on the effects upon the second hearing of knowing the content of the recording. Most commonly expressed was the notion of *orientation*: greater familiarity was represented in terms of knowing: a) what the recording covered; b) where different pieces of information fell within the recording; and c) where gaps in understanding had occurred that might need to be redressed.

Here are some typical comments:

---

(10) R: was it different listening the second time or was it the same?
S: I think it quite different before + because to to me the first time + I don't know what is it + so I need to guess a lot…so maybe some + some information I guess is wrong + and er after I + after the first time I I heard the + er radio and maybe some information I have got + and the second time it is more easier to guess *[YJa: 99-102]*

(11) S: *I think the second time listening helped me to understand better er the + this in more details… because the first time + even I can't get + can't catch the answer and er + the second time when it goes to the er + to the place where the answer will + will appear I will put more attention to the sentence* [Ea: 154-6]

(12) S: *when I am listened the second time I'm familiar with this text*
R: *right + so how did that help you?*
S: I *can certain some answers…I can listen the missing parts …*
R: *so you can remember where parts were missing*? [Hb: 177-84]

(13) S: *after the first time I will know which part should I emphasise on during the second time…*
R: *you know where to find the information… that you want*
S: *so after the the first one I will + I will divide the whole lecture into different part… I know this answer will be in this part + this small part + so I will focus on it + so I pick it out* [Oa: 267-272]

(14) S: *sometimes I when I hear from the in the first lecture I can't hear words and I'm not sure what the word is and then I listening in the second time I can more focus and maybe I listen twice so I can guess what the words means.* [Tb: 196]

In short, some participants were able to construct a framework as a result of the first play, which enabled them to retrieve additional information that they needed. Others stored mental notes of points where they had had difficulty at word or at information level and focused on those points during the second play.

The ability on a second hearing to focus attention on particular areas of the text was an aspect foregrounded by a number of speakers. Here are two further examples

---

(15) R: *why was it different?*
S: *because I am at the first time + er I feel little nervous + and er didn't enter the er listening environment quickly + so perhaps can't + I can't make a lot of information + and in the second time I was familiar with this er more + so um perhaps when I listening + I can I can pay more attention in the in this question* [YKa: 211-212]

(16) S: *the first time + you just know little about this recording and + yes and so you have to um focus your attention to many things + yes but the second time you know much more things + and your attention can focus on something that [made] you confused* [YLa: 169]

Other less commonly represented benefits of familiarity included (YG) enabling the listener to gain a wider perspective after focusing quite narrowly on answers to questions. This can be explained in terms of attentional load. Having established a rough overview of the recording, this participant found it less necessary to focus upon details of the text that might be keys.

Three participants (YB, O, Q) commented that they felt, during the second play, that they understood in greater depth the full implications of the *items* that they were attempting to answer. One (YL) commented that familiarity with the text made her feel during the second play that the speaker was speaking less rapidly than she had originally assumed.

5.2.2 Anxiety

Nineteen out of 36 participants made comments that were explicitly or indirectly related to the phenomenon of listening anxiety. The following were representative.

(17) S: *because I am at the first time + I feel little nervous and didn't enter the listening environment quickly* [YKa: 212]

(18) R: *do you think you were listening differently + when you listened the second time?*
S: *yes more relaxed + I felt more relaxed in the second time*
R: *okay + you felt nervous [during the single-play task]?*
S. *yes, yes… because if I were + if I lose the answer + I no more + I don't + I won't have any more chance*
S: *yes, so I'm er so I was a little nervous in the first time* [Aa: 98 – 105]

(19) S: *no + in the last + the second time + I feel more relaxed and I can have more detail* [Jb: 210]

Only three participants (I, O and S) said that they felt little or no difference in confidence when participating in the second play – while a fourth (T) was ambivalent, asserting that unfamiliarity with the topic was a greater threat to her confidence than hearing a recording only once.

5.2.3 Attention

Human attention is limited in capacity and the mark of a successful listener is the ability to direct attention where it is most needed. Issues of attention underlie some of the discussion in Section 5.2.1 above, where a second play apparently enabled test-takers to move away from too narrow a focus upon word level information and to free up attentional capacity that enabled them to take account of wider issues. The function of attention that we consider in this section relates to the specific demands made upon the test-taker by particular tasks.

The types of format conventionally employed in tests of L2 listening impose additional cognitive demands upon the listener that would arguably not be present in real-world circumstances. Multiple-choice questions require close reading, the balancing of conflicting options and the ability to make rapid decisions between them. Even more cognitively demanding are gap-filling tasks, where a candidate has to read and to write at the same time as absorbing information from an ongoing recording.

The divided attention effects associated with conventional comprehension tasks were referred to indirectly by a number of participants. They were mentioned in relation to the type of listening anxiety covered in Section 5.2.2, with four participants associating anxiety with the pressures imposed by test formats. Three of them (YD, L and R) commented upon the need to move quite rapidly from one question to another, and consequently welcomed the opportunity of revisiting the questions.

(20) S: *That that means like maybe if I just stay a second …+ and try to think about + I can like imagine what it could be + but …I will have no time because it's just for play once time + I think I need to move to the next point …. [With two plays] I catch it the next time because I'm sure I have the time and to think about it.* [YDa: 138]

(21) S: *when I hear the first time + I will be very nervous… because I + I am afraid that I will miss something important… and when I hear some key words + I will want to write it down + and when I write down maybe I will miss the next sentence* [La: 217-221]

(22) S: *first time I feel not confident + because that will be kind of be once + and I'm afraid of miss anything in it* [Rb: 194]

One participant (G, reporting on the gap filling task) commented upon the need to write and listen at the same time and on the time pressures imposed by a single play situation.

(23) S: *but this time the first time I just + I don't need to write the whole er + I don't worry that if I forgot the word.*
R: *you could finish writing… the second time?*
S. *yes + I can do that* [Gb: 186-190]

In addition, two further participants (YB and Q) mainly envisaged the advantages of a second hearing of a recording in terms of the way it allows test-takers the ability to listen more freely without the heavy cognitive demands imposed by test formats:

(24) S: *IELTS test is + especially the section 4 + you have little time to + read the question and you must er + focus on everything about it and follow this + follow this question + maybe you can't er + you can't suit + you can't be suitable about the first time + but you know the meaning you know all all of these questions + and for the second time you can focus on it + you have a memory and brain + yes the second time is more easy* [YBa: 88]

(25) R: *Did you find it easier listening the second time?*
S: *yes I think it's easier because I don't have + I don't need to spend some extra attentions to read the text so*
R: *… you're telling me that you know the questions better?*
S: *yes*
R: *and therefore you don't have to keep reading the questions carefully?*
S. *yes, and I can concentrate on listening* [Qb:135-139]

5.2.4 Participants' accounts of listening goals

A final area of discussion relates to learners' own perceptions of how they focused on the input during the second play; and the extent to which their targets differed from those they had sought during the first play.

First vs second play

The topic was approached from two different angles. Firstly, it was of interest to put to the test the provisional hypothesis put forward by Buck (1990) and by Field (2013) that the first play gives rise to a more focused (and perceptually based) approach which seeks keys to questions, while the second enables the listener to adopt a wider (more conceptually based) view of the recorded content and even to recognise patterns of argument in what the speaker says.

Comments from participants relating to this were generally elicited by the standard question about the 'differences' they had noted between their performance during Play 2 as against Play 1. The question was phrased in a way that was intentionally general so as to avoid biasing participant responses, so only 17 of the 36 participants responded with information that related to the relevant parameters (specific / general, local / global, test-specific / topic-specific). Of those who did so, four participants (B, E, F, I) maintained that there was no difference between the two plays and that they behaved identically in each – though two of them contradicted this claim elsewhere in their interview.

The majority of the respondents who identified a change in their behaviour (9 out of 15) described the shift as being from specific to more general. Eight of them (A, C, J, K, O, S, YA, XD) described an approach focused on seeking answers to questions during the first play which was then followed by more general listening during the second.

(26) S: *for the second time + I can pay more attention about the + the whole + the whole ++ the whole information in the record* [YAa: 34]

(27) S: *the first time my + in the first time er + I think the main aim + my main aim is just to find answer of the questions…and the second time + however in the second time I tried to understand the all + understand all content + content* [Aa: 121-123]

(28) R: *when you listened again … was it a different kind of listening?*
S: *oh different kind of listening*
R: *or was it just the same, really?*
S: *more or less will be the same + but I will maybe hear more clearly the whole + the whole information + but not just focus on er + a key word + I will try to understand the whole text* [Sb: 193-196]

Participants J and XD went further and characterised Play 1 as marked by the use of test-wise strategies and Play 2 as less constrained listening.

(29) R: *were you listening in a different way?*
S: *yes, if we listen for twice + we + were + we can just check the answer that we're not sure about + and then heard more information about the + the transcript because when we do the exam + at first time we just focus on the gap we need to fill + we + sometimes we don't focus on the content + because you know in China + er sometimes we + the teacher sometimes tells us the method to do the listening test is you + you + is fine for you do not understand about the transcript + but when you look at the key words before + before the the gap + you you know that the answer should be there*
R: *so you think the second listening…helped you to move away from that?*
S: *yes + you can maybe understand more about the + the + the transcript + not only focus on the the gap we should full + we should fill but also the information + the whole content* [Jb: 193-198]

These comments accord with what one might have predicted (and indeed with Buck's early finding). But there was by no means universal agreement. Two participants reported the opposite course: moving from general listening in the first play (orienting themselves to information in the recording and where it was to be found) to a more targeted approach in the second.

(30) S: *the first time I heard I heard the record + I will concentrate all the things + and the second time + I only have to concentrate the questions + and something related to the questions* [F: 158]

Others (E, L, R, XA) characterised the first play as being used for gathering information, while the second enabled them to extract finer details

(31) R: *so you said the first time helped you find out where the answers were*
S: *yes.*
R: *and so the second time*
S: *maybe sometimes first time I can't catch them but the second time I will hear some more details* [Ea: 157-60]

(32) *R: the second time?*
S: except the + maybe not that general idea + this whole part … but specific ideas on this. [Rb: 185-88]

(33) S: *from first I got some idea + even a little bit …and for second time I'm developing the idea.* [XAa: 165-7]

If these accounts provide reliable insights into processing (with due allowance made for participants over-reporting or over-simplifying what occurred), either approach would seem to make sense. Targeting questions at the outset ensures that provisional answers are obtained on a first pass and that any gaps or uncertainties can be clarified on a second listening, which could also extend to taking in information beyond that focused on by items. The opposite approach provides the test-taker with a conceptual framework for the text as a whole and points of reference for different sub-topics, thus enabling answer keys to be located more efficiently. Both constitute a combination of the perceptual and the conceptual. They do not solely provide conceptual information in the form of a more integrated view of recording content, as Buck (1990) concluded; they also entail a focus on perceptual information with a view to consolidating first impressions of the wording used and to checking on grey areas.

What becomes apparent is the extent to which the demands of the test – i.e. the need to find answers to questions – inform not only the first but also the second play. But this does not cancel out other evidence that a major difference in the double play format is the parallel opportunity to engage in higher-level listening processes relating to the wider meaning of the recording. Such processes form an important part of the real-world listening construct – especially for those who are operating in academic contexts.

Participants' accounts of targets

A second group of responses were given in answer to the question: What did you pay attention to when you listened for a second time? The majority of the 15 who responded unambiguously on this (E, K, O, P, S, T, U, V, YD, YG) reported a combination of words and ideas.

(38) R: *how was it different?*
S: I can catch er more more words and understand more understand
R: did you + you caught more words + did you think you understood more ideas?
S: yes [YGa: 105-108]

(39) R: *I'm trying to say + what made it easier for you to find answers*
S: *because I can catch more + more words and and + what I understand, is more er clear*
[Pa: 189-190]

(40) [R: *did you hear some new words that you hadn't heard before?…new ideas?*]
S. *new words. I think*
R: *…so you heard words…that you hadn't heard before + or maybe that you hadn't heard clearly before?*
S: *and I also had some time to rethink the understanding of the sentence* [Va: 198-204]

Tellingly, however, only one of these participants mentioned a better understanding of the line of argument of the speaker.

(41) R: *did you pick up words that you hadn't heard before?… ideas that you hadn't heard before?*
S: *maybe idea + idea + I have a more understanding about the macro + macro + how to say*
R: *well I know what you mean + a macro structure*
S: *a global picture…of the + of the lecture, yes* [Oa: 302-306]

While another was quite explicit about *not* grasping the line of argument

(42) S: *I hear more detailed information*
R: *right so you heard some of the detail +did you hear words that you hadn't heard before?*
S: *yes*
R: *do you think that you understood the + the argument + the way that the person was presenting the information?...*
S: *not much* [Sb: 164-168]

One further participant (J) asserted that he had listened at the level of ideas rather than words the second time.

(37) R: *was it ideas + or was it just words that you recognised?*
S: *ideas + the whole + maybe the main points about the the the lecture + because the first time just focus on the gaps* [Ja: 235 – 236]

But once again there were dissenting voices. Four participants (YE, G, H, Q) reported that during the second play they were chiefly aware of targeting words rather than wider meanings. The presence of Q (IELTS listening 7.5) in this group indicates that this perception is not simply a reflection of lower proficiency problems of decoding but a by-product of the testing situation.

---

(34) R: *did you understand new ideas + new facts + or was it … just words that you recognised that you hadn't heard before?*
S: *some ideas, but only a little*
R: *right + only a little + it was mainly the words that helped you*
S: *yes + but I don't care what's the idea + I just care the answer in the text* [Gb:209-214]

(35) S: *I understand more word*
R: *so you think you understood more words?*
S*: yes.*
R: *do you feel that you understood the ideas?*
S: *no* [YEa: 243-247]

The focus on words accords with the evidence from the verbal reports. The fact that it was reported by so many participants demonstrates that (even during Play 2) it served as a deliberate strategy for accessing information in the input. But it is clearly a strategy that, for many participants, operated in tandem with widening understanding of content.

6. GENERAL DISCUSSION

SUMMARY OF FINDINGS

RQ1: To what extent does a candidate's score improve with the opportunity to hear a recording twice?

The scores of the majority of the population increased as a result of a second play by between 1 and 3 points out of 10. This confirmed findings by earlier commentators. However, it would appear that the increases in scores are not as large as sometimes assumed and vary considerably between individual test-takers.

RQ2: Is level of proficiency a factor in determining whether candidates benefit from a second play in terms of an increased score?

Level of listening proficiency did not prove to be a factor determining an increased score.

a. Scores increased after Play 2 for each of three different proficiency bands.

b. They did so in a way that was roughly proportionate and that sustained and heightened the distinctions between the proficiency bands.

c. First play scores were well below the overall IELTS listening scores which participants had obtained before enrolling on their course. It was not until after the second play that they achieved scores commensurate with the grades that they had been awarded.

RQ3: Is test method a factor in the way in which a second play is handled?

Test method was found to affect both score and behaviour.

a. Candidates working with a constructed response format were assisted by a second play to a greater degree than those working with a selective response format.

b. Evidence from the protocols indicated that, on a second play, the MCQ format fostered strategies based on checking the recording for key words and paraphrases. The gap-filling format led test-takers to check provisional answers against the input, listening out for any words they had failed to supply.

RQ4: In what ways does a candidate take advantage of the second play to check, change or supplement answers?

a. A relatively small number (13%) of the changes were positive, i.e. revised from incorrect to correct as a result of the second play, while 3% were negative, i.e. revised from correct to incorrect. Nearly half the responses were initially blank or incorrect and remained blank or incorrect after the second play. These figures suggest that a double play does not make a test markedly easier.

b. A major function of the second play was to check answers. Around a third of all responses were originally correct and were confirmed as correct on a second play.

c. Evidence from verbal reports suggested that, even during the second play, test-takers continued to focus on word-level information with a view to checking answers or decoding words that had been missed or not recognised.

RQ5: Does a candidate behave differently when it is established that a second play of the recording will be available?

a. Most participants appeared to adopt a procedure which entailed locating provisional answers to questions during the first play. They then confirmed these answers were correct during the second play and/or checked for missing answers or for words that had been missed.

b. There was lower listening anxiety, with less pressure upon the test-taker to identify responses on a once-and-for-all basis during the first play.

c. Second play processing benefited from greater familiarity with the recording – and greater familiarity with the test items and their relationship to the recording.

d. During the second play, a continuing focus on word-level information was often combined with a wider perspective on the listening text and/or a clearer grasp of the speaker's ideas.

6.1 Issues of scoring

The first step in the study was to establish that playing a test recording twice does indeed lead to an increase in scores. Results from test-takers (N = 73) operating under test conditions indicated that scores increased significantly as a result of a second play. This applied to nearly two-thirds of the population, with only 27% whose scores remained unchanged. The finding confirmed similar ones by Berne (1995) and Cervantes & Gainer (1992). However, an important goal of the present study was to take this insight further by extending our understanding of the nature of the adjustments that take place during a second play and establishing whether those adjustments compromise the fairness of the scoring.

The study examined the hypothesis (based on an earlier pilot) that a participant's level of listening proficiency might influence the extent to which he/she benefited from hearing a recording twice. The logic was that a low-proficiency candidate might not be able to decode sufficient of the input to be able to make an informed choice of response, even on hearing a recording a second time. Conversely, it was assumed that a high-proficiency candidate might be able to form confident conclusions as to the correct answers on a first hearing and might therefore not need to hear the recording again. If either assumption proved to be true, it could be argued that the convention of playing a recording twice benefited a single middle-range proficiency band disproportionately, and thus reduced a test's ability to make fine scoring distinctions between candidates.

In fact, level of listening proficiency did not prove to be a factor determining the increase in scores after a second play. Scores were found to increase within each of three proficiency bands. They did so in a way that was roughly proportionate and that sustained and indeed *heightened the distinctions between the bands*. The evidence suggests strongly that, far from blurring the score differentials between candidates (for example, by unduly favouring weaker candidates), the effect of a double play is to increase those differentials.

However, two issues of concern do emerge. Firstly (see Table 3), while the actual increase in scores between the middle (5.5 to 6) and the higher (6.5 to 7.5) bands was similar, it represented a proportionately greater level of increase for the former (34.48% as against 22.81%), taking their mean score from 3.95 to 5.32. This group is in a sensitive position – at the borderline between refusal and acceptance for university study in the UK and elsewhere (see Appendix 3). A possible conclusion might be that providing a second play in a test used for academic admission gives learners the opportunity to prove what they can achieve – always assuming that a case can be made for the construct and ecological validity of playing a recording twice.

Secondly, the mean scores achieved by participants on a first play appeared to fall well below their recorded IELTS band scores for listening. It was not until after the second play that they achieved scores commensurate with the grades that they had been awarded. This suggests that many of those who achieve high overall IELTS listening grades under single play conditions do so chiefly on the basis of their performance in the less demanding conversational sections of the test (constituting 75% of the marks). Such overall grades are unlikely to be adequately predictive of a candidate's ability to perform competently when confronted with real-world lecture-style material, which is more formal in language, is expository or discursive in discourse structure and is more information-dense.

While this is first and foremost a concern for those responsible for the structure of IELTS, there are clearly wider conclusions to be drawn in relation to the effects of single play. Considering the results from the lecture-based section on which the present participants were tested, it becomes apparent that the narrow range of the scores achieved after a first play (particularly in the case of the gap-filling task) may not be sufficient to discriminate finely between candidates' levels of lecture-listening proficiency. The data suggests that a second play can make the scores obtained more sensitive – and that it does so without disproportionately favouring any particular proficiency level.

One further question that arose was whether the second play seriously *disadvantaged* any participants (for example, indecisive individuals) by making them re-think their first accurate responses. Scores and proficiency levels were checked for the eight participants whose scores went down after the second play. All of them except two had previous IELTS band scores of 5 and below. Their loss of one or two points on the second play can thus probably be put down to random guessing. Broadly speaking, most adjustments appeared to be in the direction of a score that more adequately represented the participants' previously attested proficiency.

6.2 Issues of format

Test method was found to be a factor determining how much advantage was obtained from a double play. In terms of scores, candidates working with a constructed response format were assisted by a second play to a significantly greater degree than those working with a selective format.

This confirmed an earlier finding to the same effect (Boroughs 2003). However, what the present study shows is that it is not a simple matter of one particular format allowing test-takers to exploit the opportunities offered by a second play while another does not. Although considerable care had been taken when selecting the materials to ensure a degree of parity between the two IELTS sections in terms of recordings and items, participants undertaking the gap-filling task achieved much lower scores after the first play. If there was an apparent format effect affecting the increase in scores following a second play, it derived largely from the very low base from which gap-filling participants were operating after Play 1.[4] Rather than necessarily demonstrating that one format is unfairly assisted by a double play while the other is not, it reflected the considerable cognitive and linguistic difficulty of choosing gap-filling responses on the basis of a single play.

Gap-filling, as Field (2013, p. 131) points out, is an extremely complex task entailing the distribution of attention between three different sets of processes: reading, writing and listening. Citing test-taker verbal reports from his own (2012) IELTS study, he suggests, indeed, that it is much more demanding than any real-world listening task[5].

[4] The point needs to be stressed again that the group participants arrived at their first-play responses in the belief that they would be permitted one hearing only. There was thus no question that they were refraining from formulating responses until they heard the recording again.

[5] Note-taking is not comparable because it is the listener who generates his/her own notes in a lecture context, whereas a gap-filling task entails mastering somebody else's notes and tracing their patterns of logic.

The low first-play scores evidenced here suggest that this type of constructed task is perhaps only defensible in terms of cognitive and scoring validity if the test-taker is indeed permitted a second play in order to check the diverse information sources that he/she has to command (sentence stems plus draft answers plus audio input). Similar thinking may perhaps have informed Cambridge English's decision to add a second play to a CAE gap-filling task during the 2008 revision of the exam (Murray 2007, p. 20).

The protocols also provided evidence that format influenced the way in which the second play was handled. The MCQ task largely fostered strategies based on re-visiting locations within the recording where likely key words and paraphrases had been identified. By contrast, the gap-filling format prompted test-takers to check provisional answers against the input, sometimes with a view to supplying any words that might be missing from the original answer. In short, the MCQ approach was broadly one of checking answers while the gap-filling one quite often entailed supplementing them. Given the strict rating criteria governing the wording of constructed responses, it seems entirely reasonable to give test-takers the opportunity via a double play of checking the form of words used.

A relevant example was provided by participants' responses to a set of gap-filling items at the end of Text A. Question 8 required them to supply the name of the first movie made, the key in the recording being: '*…which led to the making of The Great Train Robbery – the very first movie made'.* Several participants were unable to retrieve the complete name of the film from working memory on a first play and wrote only part (e.g. *Train Robbery*) but they succeeded in giving the correct answer after a second hearing. This is an exceptional example where the answer cue (*the very first movie made*) occurs after the information has been given. But it still serves to underline the fact that a second play performs an important function in cases where verbatim word recall is required.

The question that then arises is: Should test designers acknowledge that a second play is desirable for certain more demanding formats such as gap-filling and not for others? The problem here is that it is difficult to determine how relatively great the cognitive demands of a task have to be for a decision of this kind to be considered. The data (Table 5) indicated markedly higher scores on the MCQ task despite the fact that the groups which undertook this task were of a lower overall proficiency than those which undertook the gap-filling one. However, this does not mean that the processes entailed in answering MCQs are necessarily straightforward and/or closely representative of the processes entailed in real-world lecture listening. In fact, the demands of focusing attention as a reader on three or four options and discriminating between them while at the same time listening are considerably greater than those generally associated with real-world listening events.

6.3 Affective issues

Language anxiety is a well-established phenomenon in SLA studies, following early work by Horwitz & Young (1991). It is said to be especially associated with the use of L2 oral skills (Horwitz, Horwitz & Cope, 1986), presumably because of the time pressures under which speakers and listeners have to operate.

There is a great deal of anecdotal evidence that learners view listening as the skill about which they feel least confident – a notion confirmed in Graham's 2006 study of British learners of French. A consequence is that listening performance can be severely affected by anxiety, both within the classroom and in real-life interaction. L2 listening anxiety has been the subject of a number of research studies and reviews including papers by Vogely (1998), Elkhafaifi (2005), Mills, Pajares & Herron (2006), Chang (2008), and Kimura (2008). Kim (2000) based an entire PhD study on it, and produced and trialled a set of questionnaire items (FLLAS). As an instrument, the questionnaire suffers from the limitations associated with this type of research methodology but the items serve to mark out listening anxiety as a skill-specific phenomenon and distinct from more general forms of language anxiety associated with limited language knowledge or with personality.

Intentionally, no reference was made to stress or anxiety in framing the interview questions to participants in the present study. However, when asked to compare the experience of a double play with an earlier test involving a single play, over 50% of the respondents, freely volunteered the information that they felt less *nervous*, more *relaxed* or more *confident* when permitted a second play. Only three participants out of 36 (8.3%) reported little or no difference in their confidence level when undertaking the double play test.

From this evidence, it is apparent that the availability of a second play serves to counteract listening anxiety in many test-takers, permitting them to make the most of their listening skills by avoiding some of the stress caused by the real-time nature of the recording or by the complex demands of the task. This might seem a very persuasive reason for favouring a double play in a listening test. One should, however, not entirely exclude the reverse argument that it makes the test less representative of the anxieties of a real-life listening event and, thus, possibly less valid in predicting how confidently a test-taker would behave in the world beyond the test.

6.4 Cognitive issues: orientation, attention and processing

The interview data suggested that, for the majority of the participants, cognitive behaviour differed markedly when they were permitted a second hearing of a recording. These differences affected the first play as well as the second. As just noted, there was an overall reduction in listening anxiety. Perhaps as a consequence of this, comments during the first play included quite perceptive judgements on the likely accuracy of responses – enabling the test-taker to tag points in the recording that might need checking during the second play.

Both the verbal reports and the interview data suggested that the input was processed differently during the second play when compared to the first. In the most commonly adopted procedure, participants located provisional answers to questions during the first play, then checked them during the second, or sought missing answers or words that had been missed or misunderstood. A minority described a different procedure – using Play 1 to gain a general understanding of where information was located and Play 2 to identify item answers or add details. In short, there was strong evidence that the use made of the second play was often strategic, with the individual checking answers to confirm or replace them, or locating keys to answers not previously identified.

A major factor shaping these processes (one mentioned by two-thirds of the participants) was the greater familiarity of the material during the second play. This enabled participants to orient themselves to the recorded material on two different fronts:

Coverage: following the first play, they had an idea of:

a. the content of the recorded material
b. the propositional content of the test items and key words in them that might occur or be paraphrased in the recorded material.

Location: following the first play, they were able to locate:

a. where in the recording certain pieces of information occurred; these could then be matched against a sequential set of items
b. certain problem areas (e.g. words that may have been incorrectly understood or not understood at all, points where understanding was incomplete)
c. parts of the recording which may have given rise to a wrong or provisional answer and that needed to be checked.

This picture of what occurs during double-play listening considerably refines the findings reported by Buck (1990) who suggested that the first play was characterised by more local listening while the second was more global. It was evident that much of the processing continued to be driven by the requirements of the test: an attention focus at the level of the word and the chunk was reported by many participants, even during the second play. But it also emerged that this low-level perceptual focus was more sophisticated than was at first apparent. Word-level information was used as a memory trigger to locate information, to eliminate MCQ options and to identify sections of the recording that were imperfectly understood.

What is more, most participants reported that they operated during the second play at the level of ideas as well as words. They managed to combine a focus on previously-identified problematic points in the recording with establishing a wider perspective on the text and/or a clearer understanding of the speaker's ideas. This might appear to be a contradictory finding but, as indicated earlier, it is entirely consistent with current views of the role of attention in listening, which, in a competent listener, can be switched at will from a local perceptual focus to a wider conceptual grasp of patterns of meaning. Judging by the protocols obtained, such switching was relatively rare during the first play, where most participants were strongly constrained by test demands.

To this extent, Buck's local / global surmise is supported. It would seem that a second play potentially enables test-takers to engage in a type of listening that suffers from fewer of the local constraints associated with finding answers to questions, and approximates a little more closely to that which a listener would need to employ in a real-world presentational or broadcast context.

The point should be made that, in such a context, the ability to operate at different levels of processing is a mark of a certain level of expertise: less able L2 listeners are likely to find that their attention is so heavily engaged by processing at word or clause level that wider issues of meaning and discourse structure escape them. This goes a long way towards explaining the score data, which indicated that the use of the second play did not, as has sometimes been asserted, disproportionately favour weaker listeners, but benefited listeners at all proficiency levels. It also picks up on Buck's argument about automaticity but ironically leads one to the opposite conclusion. *It is the test-taker with limited automatic processing who lacks sufficient attention resources to achieve a wider understanding of the ideas expressed by the speaker, even during a second play. By contrast, automatic word-meaning connections in a more expert listener free up memory capacity to consider the larger picture, supporting more confident and more accurate answers to questions.*

6.5 General conclusions for cognitive validity

It would seem, then, that the message is a rather mixed one so far as cognitive validity is concerned. On the one hand, there is evidence that a second play promotes (in part) a type of listening which approximates to a real-world model. It does indeed entail checking answers, but it also permits some listeners to take a wider perspective than can be achieved in a single play format with the imperative of focusing on one-off opportunities for getting the right responses. Playing a recording twice may help weaker listeners to double-check their recognition of certain words but it arguably advantages higher-level listeners more because it allows them to construct a framework of information and logic that reinforces the answers that they have chosen. On this analysis, the double play convention should not reduce score differentials between proficiency levels – and this study has provided evidence that it does not.

On the other hand, there is still the lingering concern that hearing the same piece of input twice compromises ecological validity by creating unnatural circumstances in which these cognitive processes operate. More importantly, the protocols underline the extent to which processing during the second play continues to be driven by the demands of the test, with participants, even skilled

ones, focusing attention at word and phrase level in order to check that their first attempts at answers are correct. There is also evidence (Section 6.2) that test format shapes the way in which this information is sought. In short, the second hearing of the recording continues to be shaped by: a) the need to find answers; and b) the need to engage construct-irrelevant processes associated with conventional tasks such as MCQ and gap-filling. It would seem that one cannot push the cognitive validity argument too far.

But there is a solution – and it is a simple one. As Sherman (1997) points out, test designers can curtail the impact of test format upon behaviour by delaying the point at which items are presented to the test-taker and introducing them only after the recording has been heard once. This procedure enables them to take advantage of one of the strengths of the double-play convention – namely, the opportunity to take stock of larger-scale meaning – but to do so during the *first* play of the recording rather than the second. Because the test items have not yet been seen, this first play is untrammelled by demands that focus listener attention narrowly on word-level decoding. It becomes purely an exercise in scene setting.

There are a number of benefits to handling double play in this way.

- It provides an initial exposure to the recording which elicits processing at both local and holistic levels without the local focus fostered by test items. This assists the building of a general comprehension model of the recording that resembles more closely the type of representation that a listener constructs in real life.
- It avoids during a first play the divided-attention pressures associated with conventional test formats, such as MCQ and gap-filling.
- It prevents the test-taker from using written test items to create a set of schematic expectations even before the recording has begun.
- Cognitively, it resembles an operation familiar from study reading, where skimming enables the language user to gain an idea of what is covered in a text before moving on to depth reading.

The approach is well suited to computer-delivered testing where the screen-shot of the items can easily be delayed until after a first play. It is less easy to envisage in relation to paper-based versions of tests because of the need to provide a question sheet in advance for all sections. But there seems no reason why items could not at some point in the future be presented on screen in the form of PowerPoint slides rather than on conventional question sheets.

Sherman (1997) argued on similar lines to those put forward here that the approach would reduce a test-wise tendency to word-level processing during the first hearing and would avoid the cognitive load associated with standard formats, such as MCQ. As noted earlier, a striking finding was that higher scores were obtained in this condition when compared with three others (conventional pre-set questions before a double play, post-set questions after a double play and a constructed response summary). An obvious interpretation would be that the approach makes a listening task ‘easier’ (Buck 2001, p. 151). However, an alternative and more nuanced view would be that scores were higher because the approach made the recorded material more accessible to the listener. The first play provided an overview ahead of seeing the test items, thus enabling the test-taker to follow a general-specific information building pattern (cf. news headlines followed by in-depth reports or an introductory outline anticipating what a lecturer will cover). Sherman’s participants were able to apply their L2 listening skills in an informed way when it came to the second play – benefiting from the kind of familiarity with the text that two thirds of the participants in the present study mentioned as having assisted them during their second hearing.

This suggests a new perspective on the old argument that the double play convention serves to compensate indirectly for the lack of handouts, PowerPoint slides and paralinguistic signalling. In this 'sandwich' format, the first play equips more competent test-takers with a conceptual framework on to which they can map the questions they are subsequently asked. It provides them, in other words, with something not entirely unlike the background that visual support would supply in a real-world lecture setting.

7. CONCLUSION AND IMPLICATIONS FOR APTIS

7.1 General review

This study has reviewed the arguments for and against the double play convention in listening tests, and has shed some light upon its impact on scoring and on the types of cognitive processing which it elicits from test-takers. Some conclusions must now be drawn on whether the use of the convention should be recommended for future tests of listening, and, in particular, whether it is appropriate for the specific needs of the Aptis listening test.

The *advantages* of double play, as listed at the outset and as expanded upon by the study are as follows.

a. It compensates for the *lack of visual and contextual clues* where a listening test is dependent upon audio material.

b. It allows test-takers to *normalise to unfamiliar voices*.

c. It compensates for the *complex cognitive demands* imposed upon test-takers by many of the more standard formats.

d. Evidence presented here suggested that the type of processing in which many test-takers engage during the second play differs from that employed during the first. It is less test-focused and entails *greater awareness of wider meaning and/or of detail*.

e. Evidence also suggested that the double-play procedure reduces *listening anxiety*.

The proposal that test items should not be made available until after the first play of the recording strengthens these arguments. It enables test-takers to establish a general background, directly compensating for the lack of visual cues in audio delivery. The first play also allows test-takers to adjust to unfamiliar voices and to do so in a way that is not influenced by prior knowledge of the questions that are to be asked or by the complex demands of test formats.

The following arguments *against* the convention were mentioned during the study.

a. A second play makes the listening task easier and thus *advantages lower-level candidates.*

b. The use made of a second play *varies according to test format*.

c. In most *real-world listening events,* a listener often has only one opportunity to hear and make sense of an utterance.

d. A critical component of successful listening is the ability to *process information automatically*; hearing a recording twice means that the listening construct is not being adequately tested.

e. Playing each recording twice *increases the length of the* test unnecessarily.

In relation to (a), the evidence obtained in the study indicated that, under double play conditions, scores increased across all levels of proficiency and continued to discriminate between them. On (b), it was indeed established that behaviour varies according to test format, even during the second play but, in scoring terms, the second play actually served to increase comparability between scores from the two formats featured (namely, MCQ and gap-filling).

Proposition (c) is incontrovertible but the point was made that the test formats used in language testing impose heavy cognitive demands upon the listener that render the circumstances of a test very different from those of a real-world listening event. Proposition (d) was challenged: it was argued that the ability to derive benefit from a second play is restricted in the case of test-takers with limited automaticity of processing.

The importance accorded to proposition (e) may well depend upon the scope and goals of the individual test – although account also needs to be taken of the attention of the test-taker, which probably has an upper limit of around 45 minutes. Buck claims (2011) that the single-play format is time-efficient: more tasks can be included, thus ensuring coverage of a 'greater construct range'. This is a valid point but it can only be achieved in practice if guidelines ensure that item writers tap into a greater variety of cognitive processes than the rather limited focus on local factual information that usually predominates.

On balance, there is much to be said in favour of the use of the double play procedure and little to suggest that it compromises test accuracy. It has been represented as especially beneficial if test items are only delivered after the first play, thus providing a phase of the test which approximates more closely to real-world behaviour than does listening under the constraints of mapping from item to recording. To be sure, the single play format replicates the one-pass experience of much real-world listening but one cannot overlook the non-ecological impact of test format upon the processing that takes place or the listening anxiety generated by having only one opportunity to retrieve an answer. That said, some test providers may well continue to favour a single play because they feel that it is more time-efficient.

7.2 Recommendations concerning Aptis

The present format of the Aptis listening test was described at the beginning of this report. It is a computer-delivered test which takes advantage of its mode of delivery to allow candidates a free choice as to whether they hear a recording once or twice (British Council 2013). Tests consist of 25 short recordings with one multiple-choice question per recording pre-presented on-screen. The only time constraint is an overall limit of 50 minutes. By using a single play throughout, it is possible to complete the test in 25 minutes. However, it seems that there is no reward in terms of increased scores for finishing the test ahead of time and factoring in such a measure is not thought by the test provider to be a practical possibility.

Declining the opportunity of a second play might seem an unlikely decision for many L2 listeners to make, given their awareness of the transitory nature of the signal and of their uncertain decoding skills. On the other hand, it must be borne in mind that items in the Aptis test are designed to cover a range of proficiency measures from CEFR A1 to C1. It suits higher proficiency candidates to be able to opt out of a second play if they are certain of their responses to items targeted at lower levels.

One issue to be considered is whether the continued use of a double play can be defended in providing a rationale for the Aptis test. If it serves no purpose (e.g. if most test-takers choose to do without the second play) or distorts scoring differentials, then it should perhaps be dropped altogether. This might allow extra items to be added to the test – thus increasing the coverage of each of the five CEFR levels represented.

The findings of the present study suggest the following grounds for retaining the second play convention.

- The availability of the second play seems likely to reduce listening anxiety, whether or not advantage is taken of it.
- Double play allows test-takers to normalise to the voice or voices used in the recording – an important consideration when short recordings are used, as they are in Aptis.
- Double play compensates for the lack of visual or paralinguistic cues in what is at present an audio-only presentation.
- On evidence obtained in this study, a second play is unlikely to render the test too easy or to unduly advantage lower level students. The ability of the test to discriminate between candidates is not likely to be affected – indeed, it may be enhanced.

Because the second play is voluntary in Aptis, on-screen presentation of test items takes place before the recording begins and a between-plays presentation of items is not possible. This is in some ways a pity as a computer-delivered test like Aptis lends itself well to this procedure. It was suggested above that a major advantage of delivering test items between plays is that it encourages test-takers to create a wider representation of the content of a recording – one derived solely from oral input, on to which written test items are only mapped during a second hearing. The first play elicits listening processes which more closely resemble real-world ones in that they embrace wider meanings as well as purely local ones and are not distorted by the use of test-wise strategies. Far from unfairly benefiting weaker listeners, this version of a double play advantages more proficient listeners, whose greater automaticity of processing allows them to accord attention to the larger picture. In short, introducing this feature into the Aptis test would be defensible on the grounds of both cognitive validity and scoring validity.

A double play procedure would also accord well with any moves to extend the range of the recorded material featured in the test, in order to represent more extensively the academic and professional discourse with which at least some of the target population have to deal. Monologue presentational style material (both expository and discursive) is particularly suited to the general / specific sequence that a between-play procedure appears to support.

The disadvantage of making such a change would be the loss of the optional second play, which, as noted, suits the multi-level nature of the test well as it enables more proficient candidates to progress rapidly through material aimed at levels A1 to B1. In weighing the pros and cons of switching to a between-plays formula, test designers may wish to consider the extent to which current test-takers currently do – or do not – take advantage of the optional second play. They may also wish to bear in mind that there are other ways of ensuring that lower-level items do not slow down higher-level candidates unduly. One entails reducing the length of the recordings at levels A1 to B1, where the processes targeted mainly entail extracting simple facts at word, phrase or clause level and do not necessarily require a one minute recording. Rather longer passages than at present might be provided at levels B2 upwards to test the candidate's ability to organise and structure information; here, featuring two questions after each passage would ensure the test remained time-efficient. Another possibility would be to order the items by level (at present, they are randomised) so that test-takers are aware of progressive difficulty.

The score data and verbal reports collected as part of this study have two further, rather more peripheral, implications for the Aptis test. Firstly, they demonstrated that test format has a considerable impact both in terms of the potential scoring differential between a first and second play and in terms of the effects upon second play processing. This suggests that the Aptis team would be well advised to vary the formats employed in the test, which are at present restricted to written MCQ.

Secondly, it became clear in the study that test-takers directed a great deal of attentional focus at low-level word and phrase identification even during a second hearing. This is an inevitable consequence of the testing situation but it illustrates how many of the items in the two IELTS tests studied target pieces of local factual information signalled by word-level cues and how few tap into wider interpretations of inferred context, speaker attitude, line of argument, etc. To their credit, the Aptis test designers attempt to move beyond the purely factual, with three levels of processing difficulty characterised loosely as 'word and number recognition', 'literal meaning' and 'inference'. But the item writer guidelines should perhaps be expanded to ensure a wide range of cognitive operations which reflect both the size of the unit of information being targeted and the associated cognitive processes.

7.3 Final remarks

This would appear to be the most extensive study to date of the effects of double play upon test-taker performance. Clearly, there are certain areas that it has left unexplored. It would be interesting for a future research study to compare the performance of the same individuals operating in single play versus double play conditions. It would also be of value to extend the study beyond material like that in IELTS which was originally designed for a single play[6], and to investigate the impact of double play upon tests such as Aptis which make use of shorter pieces of recording and single items.

That said, the study has added considerably to our understanding of the effects of double play on scoring validity, the extent to which double play assists particular proficiency levels, the effects of test method upon both first and second plays, and the differences in the cognitive processes employed by learners when hearing a recording for the second time.

[6] Though it has to be said that a close examination of the material used here showed few signs of the redundancy which the IELTS examiners sometimes suggest they factor in to their scripts in order to compensate for the single play situation.

REFERENCES

Badger, R. & Yan X. (2012). The use of tactics and strategies by Chinese students in the listening component of IELTS. In L. Taylor & C. Weir (Eds.) *IELTS Collected Papers 2: Research in Reading and Listening Assessment*. Cambridge: Cambridge University Press.

Berne, J.E. (1995). How does varying pre-listening activities affect second language listening comprehension? *Hispania*, vol. 78, pp. 316–329.

Bond, Z.S. (1999). *Slips of the ear: Errors in the perception of casual conversation*. New York: Academic Press.

Boroughs, R. (2003). The change process at paper level. Paper 4: Listening. In C.J. Weir & M. Milanovich (Eds.), *Continuity and Change: Revising the Cambridge Proficiency in English Examination: 1913–2002*. Cambridge: Cambridge University Press, pp. 315–366.

British Council. (2013). *Aptis Candidate Guide July 2013*. London: British Council. Available online at www.britishcouncil.org/exam/aptis

Buck, G. (1990). *The Testing of Second Language Listening Comprehension*. Unpublished PhD dissertation, University of Lancaster.

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

Buck, G. (2011). *Testing Listening Comprehension*. North Fork, CA: Lidget Green Publishing.

Cambridge ESOL. (2008). *Cambridge Advanced English Handbook for Teachers*, 2008. Cambridge: Cambridge ESOL.

Cervantes, R. & Gainer, G. (1992). The effects of syntactic simplification and repetition on listening comprehension. *TESOL Quarterly, 26*(4), pp. 767–770.

Chang, C.A. (2008). Sources of listening anxiety in learning English as a foreign language. *Perceptual and Motor Skills, 106*, pp. 21–34.

Elkhafaifi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *The Modern Language Journal,* vol. 89, pp. 206–220.

Elliott, M. & Wilson, J. (2013). Context validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining Listening*. Cambridge: Cambridge University Press.

Ericsson, K.A. & Simon, H.A. (1993). *Protocol analysis: verbal reports on data,* 2nd edn. Cambridge, MA: MIT Press.

Field, J. (2008a). *Listening in the Language Classroom*. Cambridge: Cambridge University Press.

Field, J. (2008b). Revising segmentation hypotheses in first and second language listening. *System*, vol 36, pp. 35–51.

Field, J. (2009). Two bites of the cherry: the effects of replay on the listener. Paper presented at the BAAL Annual Meeting 2009, Newcastle.

Field, J. (2012). The cognitive validity of the lecture listening section of the IELTS listening paper. In L. Taylor & C. Weir (Eds.), *IELTS Collected Papers 2: Research in Reading and Listening Assessment*. Cambridge: Cambridge University Press.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining Listening*. Cambridge: Cambridge University Press.

Fortune, A.J. (2004). *Teaching listening comprehension in a foreign language – does the number of times a text is heard affect performance?* Unpublished MA dissertation, University of Bristol, UK.

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

Geranpayeh, A. & Taylor L. (2008). Examining listening developments and issues in assessing second language listening. *Cambridge ESOL Research Notes*, 32, pp. 2–5.

Glaser, R. (1991). Expertise and assessment. In M.C. Whittrock & E.L. Baker (Eds.), *Testing and Cognition*. Englewood Cliffs: Prentice Hall, pp. 17–30.

Grosjean, F. (1985). The recognition of words after their acoustic offsets: Evidence and implications. *Perception and Psychophysics*, vol. 38, pp. 299–310.

Horwitz, E.K. & Young, D.J. (1991). *Language anxiety: From theory and research to classroom implications.* Upper Saddle River, NJ: Prentice Hall.

Horwitz, E.K., Horwitz, M.B. & Cope, J. (1986). Foreign language classroom anxiety. *Modern Language Journal,* vol 70, pp. 125–132.

Kim, J. H. (2000). *Foreign language listening anxiety: A study of Korean students learning Engl*ish. Unpublished doctoral dissertation, University of Texas, Austin.

Kimura, H. (2008). Foreign language listening anxiety: Its dimensionality and group differences. *JALT Journal, 30*, pp. 173–196.

McGurk, H. & McDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, pp. 746–748.

McQueen, J. (2007). Eight questions about spoken word recognition. In G. Gaskell (Ed.) *The Oxford Handbook of Psycholinguistics*. Oxford: Oxford University Press, pp. 37–54.

Mills, N., Pajares, F., & Herron, C. (2006). A reevaluation of the role of anxiety: Self-efficacy, anxiety, and their relation to reading and listening proficiency. *Foreign Language Annals, 39*, pp. 276–295.

Murray, S. (2007). Reviewing the CAE listening test. *Cambridge ESOL Research Reports*, 30, pp. 1–28

Pisoni, D.B. (1997). Some thoughts on normalisation in speech perception. In K. Johnson & J.W. Mullenix (Eds.), *Talker Variability in Speech Processing*. San Diego: Academic Press, pp. 33–66.

Sherman, J. (1997). The effect of question preview in listening comprehension tests. *Language Testing* 1997, 14, p. 185.

Vogely, A.J. (1998). Listening comprehension anxiety: Students' reported sources and solutions. *Foreign Language Annals, 31*, pp. 67–80.

Weir, C. (2005). *Language Testing and Validation: an Evidence-based Approach.* Basingstoke, Palgrave Macmillan.

Appendix 1:

Materials used in group and individual tasks

TEXT A [*IELTS 6* (2007) Test 2, Section 4]

I confirm that I am willing to do this task as part of a research study into how students listen during second language listening tests. I understand that the results will be anonymous and that my name will not appear in them.

Signed ……

Name …….. Date……..

The history of moving pictures (former IELTS test)

Circle the correct letter, A, B or C

1 Some photographs of a horse running showed
- A all feet off the ground
- B at least one foot on the ground.
- C two feet off the ground A B C

2 The Scotsman employed by Edison
- A designed a system to use the technology Edison had invented
- B used available technology to make a new system
- C was already an expert in motion picture technology A B C

3. One major problem with the first system was that
- A only one person could be filmed
- B people could only see very short films
- C the camera was very heavy A B C

4 Rival systems started to appear in Europe after people had
- A been told about the American system
- B seen the American system
- C used the American system A B C

5. In 1895, a famous new system was developed by
- A a French team working alone
- B a French and German team working together
- C a German team, who invented the word 'cinema' A B C

6 Longer films were not made at the time because of problems involving
- A the subject matter
- B the camera
- C the film projector A B C

7. The 'Lantham Loop' invention relied on
- A removing tension between the film reels
- B adding three more film reels to the system
- C making one of the film reels more effective A B C

*Complete the sentences below. Write **NO MORE THAN THREE WORDS** for each answer.*

8 The first motion picture was called *The* ……

9 …….. were used for the first time on film in 1926.

10 Subtitles were added to *The Lights of New York* because of its ……

RECORDING A: TAPESCRIPT

Many believe that the story first began in America in 1877, when two friends were arguing about whether a horse ever had all four feet or hooves off the ground when it galloped. To settle the bet, a photographer was asked to photograph a horse galloping and the bet was settled, because you could see that all the hooves were off the ground in some of the photos. What was even more interesting was that if the photos were shown in quick succession the horse looked like it was running – in other words 'moving pictures'.

The person who became interested in taking the moving pictures to its next step was the famous American inventor Thomas Edison. Actually, he didn't do the work himself but rather asked a young Scotsmen in his employ to design a system, which he did. Now this young fellow was clever because the first thing he did was study other systems – primitive as they were – of moving pictures and then put all the existing technologies together to make the first entire motion picture system. He designed a camera, a projection device and the film. The system was first shown in New York in 1894 and was really very popular. Apparently, people lined up around the block to see the wonderful new invention. There were, however, a couple of problems with the system. The camera weighed over 200 kg and only one person at a time could see the film.

[PAUSE INSERTED IN INDIVIDUAL CONDITION]

Well, now, news of the new system in America travelled fast and a number of rival European systems started to appear once people had heard about it. The single problem with all the systems was they couldn't really project the film onto a screen – you know, so more than one person could see it. Then, in 1895, three systems were all developed, more or less at the same time, and independently of each other. I guess the most famous of these was by the Lumière brothers from France, and they called their system the *cinematographe* which of course is where the word cinema comes from. There were also two brothers in Germany who developed a successful system and they called it a *bioskop*.

Well, now, once the problems of projection had been solved, the next challenge for the inventors was to make the films longer and more interesting. A continuing problem at the time was that the films had a tendency to break when they were being played – a problem which was caused by the tension between the two wheels, or 'reels' as they are called, which hold the film. Now this problem was solved by two American brothers. They developed the Lantham Loop, which was the simple addition of a third reel between the two main reels, and this took all the tension away with the result that the film stopped snapping.

[PAUSE INSERTED IN INDIVIDUAL CONDITION]

So now there was a real possibility of having films of more than two or three minutes, and this led to the making of *The Great Train Robbery* – the very first movie made. It only lasted 11 minutes but was an absolute sensation, and there were cases of people watching the movie and actually fainting when the character fired a gun at the camera! Almost overnight movies became a craze and by 1905 people in America were lining up to see movies in store theatres, as they were called then.

I guess the next big step in terms of development of technology was to have people actually talking on the film, and the first step towards this was in 1926, when sound effects were first used on a film. It wasn't until the following year, however, that the first talkie, as they were called then, was made. This film featured actors speaking only during parts of the film and was called *The Jazz Singer* and it wasn't until 1928, that the first all-talking film was produced, and this was called *The Lights of New York*. Unfortunately, the sound on this early film was not very good and I believe they put subtitles on the film – that is, they printed the dialogue along the bottom of the film to compensate for this poor sound quality. Now, with the addition of sound, moving pictures became far more difficult to make…

TEXT B

[*IELTS 6* (2007) Test 1, Section 4]

I confirm that I am willing to do this task as part of a research study into how students listen during second language listening tests. I understand that the results will be anonymous and that my name will not appear in them.

Signed ……

Name …….. Date……..

Social history of the East End of London (former IELTS test)
*Write **NO MORE THAN TWO WORDS** for each answer*

Period	Situation
1st – 4th centuries	Produce from the area was used to (1) ……. the people of London
5th – 10th centuries	New technology allowed the production of goods made of (2) …… and …….
11th century	Lack of (3) …… in the East End encouraged the growth of businesses
16th century	Construction of facilities for the building of (4) …….. stimulated international trade Agricultural workers came from other parts of (5) …….. to look for work
17th century	Marshes were drained to provide land that could be (6)……on
19th century	Inhabitants lived in conditions of great (7) …….. with very poor sanitation

*Choose **THREE** letters A – G.*
Which **THREE** of the following problems are mentioned in connection with 20th century housing in the East End? Circle the letters

A unsympathetic landlords
B unclean water
C heating problems
D high rents
E overcrowding
F poor standards of building
G houses catching fire

RECORDING B TAPESCRIPT

In the last few weeks, we've been looking at various aspects of the social history of London, and this morning we're continuing with a look at life in the area called the East End. I'll start with a brief history of the district, and then focus on life in the first half of the 20th century.

Back in the first to the fourth centuries A.D., when the Romans controlled England, London grew into a town of 45,000 people and what's now the East End – the area by the River Thames, and along the road heading north-east from London to the coast – consisted of farmland with crops and livestock which helped to feed that population.

The Romans left in 410, at the beginning of the fifth century, and from then onwards the country suffered a series of invasions by tribes from present-day Germany and Denmark, the Angles, Saxons and Jutes, many of whom settled in the East End. The technology they introduced meant that metal and leather goods were produced there for the first time. And as the East End was by the river, ships could transport goods between there and foreign markets.

In the 11th century, in 1066, to be precise, the Normans conquered England, and during the next few centuries, London became one of the most powerful and prosperous cities in Europe. The East End benefited from this, and because there were fewer restrictions there than in the city itself, plenty of newcomers settled there from abroad, bringing their skills as workers, merchants or money lenders during the next few hundred years.

[PAUSE INSERTED IN INDIVIDUAL CONDITION]

In the 16th century, the first dock was dug where ships were constructed, eventually making the East End the focus of massive international trade. And in the late 16th century, when much of the rest of England was suffering economically, a lot of agricultural workers came to the East End to look for alternative work.

In the 17th century, the East End was still a series of separate, semi-rural settlements. There was a shortage of accommodation, so marshland was drained and built on to house the large numbers of people now living there.

By the 19th century, London was the busiest port in the world, and this became the main source of employment in the East End. Those who could afford to live in more pleasant surroundings moved out, and the area became one where the vast majority of people lived in extreme poverty, and suffered from appalling sanitary conditions.

[PAUSE INSERTED IN INDIVIDUAL CONDITION]

This brief outline takes us to the beginning of the 20th century, and now we'll turn to housing. At the beginning of the century, living conditions for the majority of working people in East London were very basic indeed. Houses were crowded closely together and usually badly built, because there was no regulation. But the poor and needy were attracted by the possibility of work, and they had to be housed. It was the availability, rather than the condition, of the housing that was the major concern for tenants and landlords alike.

Few houses had electricity at this time, so other sources of power were used, like coal for the fires which heated perhaps just one room. Of course, the smoke from these contributed a great deal to the air-pollution for which London used to be famous. A tiny, damp, unhealthy house like this might well be occupied by two full families, possibly including several children, grandparents, aunts and uncles. Now, before I go on to health implications of this way of life, I'll say something about food and nutrition…

Appendix 2:

Documents for individual participants

University of Reading

International Study and Language Centre
HumSS Building
The University of Reading
Whiteknights, PO Box 241
Reading RG6 6AA

Researcher:
Dr John Field
Phone: 0208-348-7597
Email: j.c.field@reading.ac.uk

INFORMATION SHEET

This research project investigates the way in which candidates handle listening passages and different types of question in listening papers such as those of the IELTS test. Its results will be transmitted to the British Council, which is funding the research, and they may be published by them.

You have been chosen to participate in this project on the basis of your first language, your IELTS listening score and the fact that you have not lived in the UK. You will be asked to listen to two recordings of lectures taken from previous IELTS papers, and to answer questions on them. The recordings will be paused from time to time and the researcher will ask you to give your answers and to say why you chose them. Your comments will be recorded by the researcher. The entire session will last about 35 minutes.

The recorded data will be stored on the researcher's own laptop and then transferred on to CD. All recordings will be anonymous and identified only by a letter of the alphabet. After transcriptions have been made from the CDs, the laptop recordings will be erased and the CD recordings will be stored in a locked filing cabinet in the researcher's home.

If at any point you do not wish to continue with the project, you may leave it. You will be paid ten UK pounds (£10) for your participation.

This project has been subject to ethical review by the University Ethics and Research Committee, and has been allowed to proceed.

Consent Form

Project title: **Number of plays in IELTS tests of second language listening**

I have read and had explained to me by Dr John Field the Information Sheet relating to this project.

I have had explained to me the purposes of the project and what will be required of me, and any questions have been answered to my satisfaction. I agree to the arrangements described in the Information Sheet in so far as they relate to my participation.

I understand that my participation is entirely voluntary and that I have the right to withdraw from the project any time.

I have received a copy of this Consent Form and of the accompanying Information Sheet.

Appendix 3:

Minimal IELTS admission criteria across a random range of universities (Academic Year 2014–15)

University	Undergrad: minimum IELTS score (minimum in each paper)	Postgrad: Minimum IELTS score (minimum in each paper)
Cambridge	7.5 (7.0)	7.5 (7.0)
Manchester	6.0* (5.5)	6.0-7.5
Leeds	6.0* (5.5)	6.5 (6.0)
Swansea	6.0	6.5
Brighton	6.0	6.5
Bedfordshire	6.0 (5.5)	6.0 (5.5)
Reading	6.5	6.5
Newcastle	6.5*	6.5
New South Wales	6.5* (6.0)	6.5* (6.0)
British Columbia	6.5 (6.0)	6.5 (6.0)

* Applies to most subject areas, but not all

Appendix 4:

Sample transcript of verbal report and interview

Participant code: La **Participant IELTS score: 6.5**

1	R: right + I'm going to do what we did before and ask you + we're halfway through + ask you er for your answers
2	S: uh huh
3	R: what answer did you give for number one?
4	S: er I choose A
5	R: and why did you choose A?
6	S: er because er I heard two friends are de + are betting about whether a horse when he's running all feets is + all hooves is off the ground + and they invited a photograph and they see the photographs is + actually four hooves off the ground
7	R: right + four hooves off the ground yeah + so did you hear all feet off the ground or all hooves or something off the ground?
8	S: all four hooves off the ground
9	R: hooves?
10	S: yeah
11	R: do you know what that is?
12	S: er that means feet
13	R: it means
14	S: yeah
15	R: it means a horse's feet + you learned that from the recording did you?
16	S: er yes
17	R: ok + good + well done + number two + um + what answer did you give?
18	S: er maybe it's A
19	R: you think it's A
20	S: yeah I think it's A
21	R: so you think that Edison invented this system and the Scotsman + is that
22	S: oh
23	R: Edison invented the technology and the Scotsman designed the system
24	S: yeah + yes
25	R: ok ++ what about number three? + or was it + was it something that you heard that made you think that fit number two?
26	S: um + because er in the article it mentioned Edison not development developed er moving pictures + he just have the theory about it
27	R: you think he had the theory
28	S: um maybe it's theory
29	R: mmm hmm
30	S: I'm not sure
31	R: and the Scotsman used the theory
32	S: yes his students used his theory to make the new system
33	R: tell me about number three
34	S: er I choose A
35	R: right + and why did you choose A?
36	S: er actually I hear two answers but I don't know which is the major problem + firstly it said camera was very heavy and it weighs about maybe two hundred kilos? Yeah

37 R: yeah
38 S: and then he mentioned only one person could see the film at a time
39 R: right
40 S: yeah ++ and so I choose A + I think maybe A is much more important
41 R: + um ok + but would you like to read A? + have + have a look at A
42 S: ++ look at what?
43 R: well
44 S: uh huh?
45 R: read A again
46 S: oh read A + ok
47 R: to be sure that it's right
48 S: *(mutters)*
49 R: ok?
50 S: oh
51 R: we'll we'll stay with A for the moment yeah?
52 S: uh huh
53 R: um + but you're going to have a chance to listen to it again + ok?
54 S: ok

---

55 R: right + could you tell me what your answer is to number four then?
56 S: + er it's + er it's C
57 R: it's C?
58 S: yeah
59 R: so people in Europe used the American system
60 S: mmm hmm
61 R: and then rival systems started to appear + anything that you heard that made you choose C?
62 S: *(laughs)* I just heard the words American but I don't know + heard about any other things yeah
63 R: you heard American
64 S: yeah
65 R: I presume you heard Europe as well did you?
66 S: yeah
67 R: so it was a little bit of a guess was it?
68 S: yeah maybe + it's about guess yeah
69 R: fine + number five?
70 S: um I choose C
71 R: + so you choose C?
72 S: yeah
73 R: why did you choose C?
74 S: um I heard about two German brothers
75 R: mmm hmm
76 S: so I think maybe it means a German team and I don't heard about a French
77 R: right ok
78 S: uh huh
79 R: so you didn't + you didn't think you heard anything about a French team but you thought you'd heard something about a German
80 S: German yeah
81 R: team
82 S: yeah
83 R: ok + number six?
84 S: er I choose A

85 R: + right + so the problems involved the subject matter + um anything you heard that
86 S: um + it talk about reems maybe um and the tension and reems I think maybe it's about subject matters + something like subject matters
87 R: do you know what + what subject matters means?
88 S: um + maybe it means about a + er + the the + the limity of this product or the limity of these things
89 R: no subject matter is is really what the film is about
90 S: oh + so what the film is about *(mutters)* ++ mmm maybe I want to hear it again
91 R: you might want to hear it again
92 S: yeah
93 R: ok? + but at the moment you're going to stick with A + yeah?
94 S: uh huh yes stick with A
95 R: what about number seven?
96 S: er number seven I choose B
97 R: B?
98 S: yeah
99 R: and why did you choose B?
100 S: because I hear three rooms
101 R: you hear three?
102 S: yeah I hear three
103 R: three reels?
104 S: three reels + yes
105 R: um and um + so you thought they added three reels to the system
106 S: yes
107 R: yeah?
108 S: yeah
109 R: ok + thank you for that

---

110 R: right + I'd like you now to um + we'll go through your answers again + ok?
111 S: uh huh
112 R: um could we have a look first of all at number one? are you happy with that?
113 S: er yes + I still choose A
114 R: right + what did you do? did you listen to see if it was correct or
115 S: er
116 R: you were so sure that you didn't need to listen
117 S: er I listened all four feet and hooves this time yeah
118 R: you you heard all four feet + all four hooves + yeah
119 S: yeah
120 R: ok + what about number two?
121 S: er I choose B
122 R: er does that mean you've changed it?
123 S: yeah I change it
124 R: why did you change it?
125 S: um because this time I heard the student er he used some existing technnologe
126 R: existing technology
127 S: yeah
128 R: yeah
129 S: um + he used + mmm and he used existing technologe and then to make a new + new one
130 R: right
131 S: yeah
132 R: ok + so you heard a little bit more this time than you did last time
133 S: yes
134 R: you didn't realise that maybe that was important
135 S: uh huh

136 R: the first time yeah? number three?
137 S: oh I still choose A
138 R: you still choose A
139 S: yeah
140 R: only one person could be filmed
141 S: yes + I think it is the same meaning as one person could see the film at a time + it is so
142 R: is it the same meaning?
143 S: er + it the same means + only one person can see the film + others cannot
144 R: that's what they said on the + in the recording
145 S: uh huh ++ what did they say er + er + from then on they developed another things and then more people could see the film
146 R: they did say that + they did say that at another point + yeah
147 S: uh huh
148 R: anyway never mind + number four?
149 S: uh huh
150 R: um + which one did you choose there? + did you stick with the one that you kept? + that that you chose before?
151 S: yes + I still missed this point again
152 R: you missed the point?
153 S: yes
154 R: so you still didn't didn't didn't get it
155 S: I I didn't catch any key words or something similar with this sentence or answers
156 R: right + ok
157 S: yeah
158 R: what about number five?
159 S: er number five I choose A
160 R: um and why did you choose A? because you chose another one before didn't you?
161 S: er this time I hear a French brother + a French
162 R: French brothers?
163 S: yeah French brothers yeah
164 R: yeah
165 S: and + er
166 R: so you heard the word French did you?
167 S: yeah and they + maybe the French brothers say why the word cinema
168 R: so the French brothers
169 S: yeah and then another German brothers also invented this system
170 R: right
171 S: yeah
172 R: + ok + and so you decided that B and C weren't
173 S: er
174 R: right
175 S: B is not right because in the article they don't mention that they worked together
176 R: ok
177 S: yeah
178 R: and number six?
179 S: er *(laughs)* I still choose A
180 R: you still choose A + you stick to
181 S: yeah
182 R: the subject matter + um so there was nothing more that helped you
183 S: um
184 R: in terms of number six
185 S: yes because I I still don't catch the key words I think
186 R: right + ok + number seven?
187 S: + er I stick to B

188 R: you stick to B
189 S: yes
190 R: because + any words that you heard that made you made you think that B was right?
191 S: yeah because there is a very long sentence explain how + what is this loop about + what is this loop
192 R: mmm hmm
193 S: I only catch three more rooms + these key words
194 R: right
195 S: maybe he mentioned about tensions or something else but I don't really know all the sent + the meaning of all the sentences
196 R: you didn't hear the
197 S: yeah
198 R: you didn't hear the whole so you
199 S: yeah
200 R: you didn't understand the whole sentence
201 S: yeah + I didn't understand yeah
202 R: ok + fine + thanks very much for that + now I'm going to ask you some general questions
203 S: ok
204 R: just to finish
205 S: mmm hmm
206 R: um + first of all um + did you feel more comfortable
207 S: mmm hmm
208 R: when you were able to listen to the recording twice?
209 S: er absolutely
210 R: you did?
211 S: yeah because I can confirm the information which I miss in the first time
212 R: mmm hmm
213 S: um I can make sure whether the answers I choose is right
214 R: when you listened the first time
215 S: mmm hmm
216 R: did it feel different from when you listened to the one about Antarctica
217 S: er yes because when I hear the first time I will be very nervous
218 R: mmm
219 S: because I + I am afraid that I will miss something important
220 R: right
221 S: and when I hear some key words um I will want to write it down + and when I write down maybe I will miss the next sentence
222 R: oh that's true
223 S: yeah + so if only I hear once um + I don't know whether I should write the information + important information down or just listen to all article
224 R: mmm hmm
225 S: yeah
226 R: right + so with this second one where you heard it twice
227 S: mmm hmm
228 R: did you do different things the second time from the first time? what did you do the second time?
229 S: er + maybe I will make sure some sentences I didn't hear in the first time + I will catch another key words
230 R: right
231 S: yeah
232 R: so you think you heard more words
233 S: yeah I can
234 R: do you think you heard more information as well?
235 S: um + + maybe + it is yes + more information yes

236 R: was some of the information clearer?
237 S: + um maybe more key words yeah because
238 R: right it's mainly the words
239 S: yeah mainly
240 R: that you thought were different
241 S: because if the sentences are too long I still can't understand it
242 R: right
243 S: so I only can catch some key words about these sentences
244 R: so the second time it gave you the chance to hear some words
245 S: yeah
246 R: that you hadn't heard very clearly
247 S: yes
248 R: the first time
249 S: mmm hmm
250 R: ok + um ++ any + anything that you + where you felt that you had misunderstood some information the first time and you understood it better the second time?
251 S: mmm ++ maybe it is not
252 R: no that that didn't happen
253 S: yeah
254 R: you got a clear idea the first time
255 S: yeah
256 R: the topics were did you?
257 S: yeah yes + because some details er maybe involve in the whole sentences I + and I must understand what he is talking about and then I can choose the right answer
258 R: so you think probably the first time you worked out what the person was talking about
259 S: yes
260 R: and the second time you listened to the details + especially the words yes?
261 S: yes
262 R: ok fine and um + was it easier to check the questions the second time because you knew where the information was in the recording?
263 S: er sure + yes
264 R: so that helped you a little bit because you had some idea of the
265 S: yes because I know what is this answer is and so I can + concentrate
266 R: right
267 S: on that sentences yeah
268 R: ok + thank you very much indeed
269 S: you're welcome

British Council Assessment Research Awards and Grants

If you're involved or work in research into assessment, then the British Council Assessment Research Awards and Grants might interest you.

These awards recognise achievement and innovation within the field of language assessment and form part of the British Council's extensive support of research activities across the world.

THE EFFECTS OF SINGLE AND DOUBLE PLAY UPON LISTENING TEST OUTCOMES AND COGNITIVE PROCESSING

AR-G/2015/003

John Field

ARAGs RESEARCH REPORTS ONLINE

ISSN 2057-5203