Assessor decision-making while marking a note-taking listening test: The case of the OET

Luke Harding
Lancaster University

John Pill
The University of Melbourne

Kerry Ryan
Institute for Social Research, Swinburne University of Technology

ABSTRACT This paper investigates assessor decision-making when using and applying a marking guide for a note-taking task in a specific purpose English language listening test. In contexts where note-taking items are used, a marking guide is intended to stipulate what kind of response should be accepted as evidence of the ability under test. However, there remains some scope for assessors to apply their own interpretations of the construct in judging responses that fall outside the information provided in a marking guide. From a content analysis of data collected in a stimulated recall group discussion, a taxonomy of the types of decisions made by assessors is derived and the bases on which assessors make such decisions are discussed. The present study is therefore a departure point for further investigations into how assessor decision-making processes while marking open-ended items might be improved.

1. Introduction

While raters' decision-making processes have been investigated in the context of tests of the productive skills (e.g., Brown, 2000; Cumming, Kantor, & Power, 2002; May, 2006; Milanovic, Saville, & Shen, 1996; Orr, 2002), little research has been carried out on the behaviour of raters applying marking schemes in the assessment of listening. This study was designed to investigate the nature of decision-making among human assessors on open-ended task types in a specific purpose English language test. It forms part of a broader project in which the authors have attempted to 'track' a marking guide for a note-taking listening test from its inception to its eventual use among assessors. A recent study (Harding & Ryan, 2009) examined markers' experiences while assessing Part B (presentation) of the Occupational English Test (OET) listening sub-test; the current study replicates this for Part A (consultation) of the same test.

The OET – a test of English language proficiency for qualified medical and health professionals who wish to practise in an English-language context – was originally developed

by McNamara (see 1990a). The test aims to ensure that candidates are prepared, in language terms, for the world of work in the professions of dentistry, medicine, nursing, pharmacy, physiotherapy, dietetics, occupational therapy, optometry, podiatry, radiography, speech pathology, and veterinary science. Historically, through its strong links with the Language Testing Research Centre at the University of Melbourne, the OET has been the focus of a number of widely cited empirical articles and theoretical papers which have emerged from within the community of Australian language testing scholars (e.g., Lumley & Brown, 1995; Lumley, Lynch, & McNamara, 1994; Lumley & McNamara, 1995; McNamara, 1990b, 1991; McNamara & Lumley, 1997). As a result, this Australian-based test is known in the international language testing community both as a clear example of a specific purpose test of English (see Davies, 2001; Douglas, 2000), and also as an innovative performance test (see McNamara, 1996).

The OET has been used throughout Australia since the early 1990s, and is recognized by a range of accrediting boards and councils for the health professions in Australia, New Zealand and Singapore. The OET is now administered in more than 30 countries with a candidature of over 10,000 in 2009. As part of an ongoing validation program, the OET is the subject of several current projects which have been designed to fill gaps in validation research concerning the writing sub-test (Knoch, 2009), the reading sub-test (Elder, Harding, & Knoch, 2009), the speaking sub-test (Ryan, 2007), and the listening sub-test (Harding & Ryan, 2009). The study presented in this paper forms part of this research effort through its focus on the assessment of the listening sub-test.

1.1 The OET listening sub-test

The OET is made up of four sub-tests, defined simply as the four macro skills: speaking, writing, reading and listening. The listening sub-test, which was the focus of this study, is made up of two parts, Part A and Part B, and although the organization of the test has changed over the past 20 years, the broad specifications of current versions of the listening sub-test are still essentially the same as those described by McNamara (1990a). The input for Part A is a simulated consultation between a patient and a medical professional, while in Part B candidates hear a presentation by a single speaker on a medical- or health-related issue similar in style to a professional development seminar. While each consultation and talk focuses on one main topic, a wide variety of topics is covered in the various test versions. The range of topics is aimed at preventing any particular advantage being created for one professional group among the candidates taking the test (e.g., medicine). Speakers are genuine health professionals and actors (in the case of patients) from various age groups. Authenticity is a key aspect of the input, so speakers' natural speech rates are generally deemed acceptable, although speakers might be given guidance in the recording studio. The accents of speakers have, to date, usually been native-speaker varieties (Australian, British and New Zealand English), although highly intelligible non-native speakers are not precluded from involvement. The recordings include the range of formal and less formal interactional styles a medical professional is likely to encounter in the contexts described.

The task types for each part of the test are designed to be similar to the tasks a medical professional would undertake in a parallel workplace context. The format used for the consultation in Part A represents a detailed set of notes taken by the professional during the interaction. Headings are provided to guide the candidates (e.g., Family history and

social background, Doctor's advice and patient's response), and the number of lines provided for notes and the inclusion of the maximum number of marks available for each section indicate how much information is required. Candidates are invited to include as much information relevant to the heading as they can. Tasks used for the talk in Part B include completing lecture notes, tables and charts by filling gaps or finishing sentence stems, as well as providing short answers to questions and responses to multiple-choice questions. The broad specifications for both parts are shown in Table 1.

[INSERT TABLE 1 ABOUT HERE]

During the marking period, both parts of the test are marked by 6-10 experienced assessors who are specially trained for each administration (this is because a different set of test materials is used for each administration of the test). Scoring is achieved with reference to a detailed marking guide developed at the Language Testing Research Centre. Both parts of the test work consistently well, and regularly achieve reliability estimates of around 0.9 separately, and over 0.9 together (calculated with Cronbach's alpha). Raw scores from the two parts of the test are weighted equally to give the final score.

One perennial challenge of both styles of listening assessment (note-taking in Part A, and a range of open-ended tasks in Part B) is that these types of open-ended tasks permit an element of subjectivity in the marking process (Alderson, 2000; Bachman & Palmer, 1996, 2010; Weir, 1993). Although a detailed marking guide is developed to stipulate precisely what kind of response test designers accept as evidence of the ability under test, there remains some scope for assessors to apply their own interpretations of the construct in judging responses that fall outside what is provided in the marking guide. With these concerns in mind, this study was designed to investigate the nature of decision-making in marking the note-taking (Part A) component of the OET listening sub-test.

1.2 Assessing notes in Part A

In order to illustrate what assessors routinely encounter while assessing notes on the OET, an example section from Part A and its related marking guide entry are given in the figures below. Figure 1 shows an exemplar task where candidates are required to take notes on an excerpt from a conversation between a patient (Clare) and her optometrist, in which Clare's contact lenses are discussed together with a particular problem she is experiencing. The corresponding entry in the marking guide is shown in Figure 2.

[INSERT FIGURE 1 ABOUT HERE]

[INSERT FIGURE 2 ABOUT HERE]

The conventions of the marking guide are that bracketed information is optional, while a forward slash '/' indicates that either adjacent word or phrase is acceptable (i.e., 'problems recent' or 'problems in last year' are both acceptable responses for the same mark). Other conventions not illustrated in this extract are described in Author (2009).

While the marking guide is designed to be as accurate and comprehensive as possible, its contents do not comprise a list of the only acceptable answers. This is primarily because the importance of task authenticity in the construct requires that a certain level of

flexibility be allowed in judging the correctness of the information conveyed in a candidate's notes, as note-taking is an idiosyncratic exercise, and information may be recorded in a variety of ways (see McNamara, 1996, pp. 109-110). Specifically, the assessors for the OET are instructed to accept spelling variations and misspellings; to accept abbreviations (which are used extensively within the health professions); to disregard grammatical errors that do not affect meaning; and to consider synonyms or alternate phrasings of answers on their merits. The guidelines that assessors read on their marking guides instruct:

> The essential point to keep in mind is whether an answer indicates an appropriate response to the question, not whether it follows the suggested answer verbatim.

Thus, for the candidate's response shown in Figure 1, as well as marks for (3b) and (3d), the candidate would have received a mark for (3g) because the term 'trouble' is a clear synonym for 'problem'.

1.3 The potential for assessor variability

While these flexible marking parameters arguably allow for a scoring procedure which captures a candidate's ability to listen more accurately than if the key was rigid, the marking guidelines also pose challenges both for test developers and for assessors when candidates' responses are expressed in a manner which does not directly match the information given in the marking guide, which is inevitably limited for reasons of practicality. In these cases, the onus falls upon the assessor to interpret the correctness of a response, a process which Buck (2001) describes as potentially problematic in two ways: (1) 'determining what constitutes a reasonable interpretation of the text', and (2) determining 'what constitutes a sufficient response to the question' (p. 140). This particular disadvantage of using open-ended task types in the assessment of comprehension has been noted in several texts (e.g., Alderson, 2000; Bachman & Palmer, 1996, 2010; Weir, 1993), and it is recommended that marking guides should be as comprehensive as possible, and should ideally be trialed and refined through an iterative development process to include a variety of acceptable responses. The scope for interpretation is further 'managed' within the assessment procedures for the OET through a markers' meeting at the beginning of each marking period. The meeting provides a forum in which assessors can become acquainted with the guide, and with some examples of responses which may potentially be more difficult to evaluate; however, it cannot prevent unanticipated responses arising in candidates' scripts during routine marking which present novel problems and require assessors to make on-the-spot decisions.

Following the argument-based approach to validation outlined by Kane (1992, 2006), the presence of subjectivity in the marking process may be conceptualized as potentially weakening the interpretive claims of the test. If assessors need to make decisions about responses, and if these decisions are not consistent across assessors, then this variability would comprise a source of error in the test scores (Xi, 2008). This problem of rater bias, which is inherent to any scoring system which contains subjective decisions, has been widely researched in the testing literature with a focus on the rating of speaking and writing tasks (e.g., Bachman, Lynch, & Mason, 1995; Eckes, 2008). Within this broader field of rater behaviour research, several studies have attempted to describe the decision-making processes which assessors employ as they carry out the rating task, observing that raters

may make different decisions depending on their experience and background characteristics (see Cumming, Kantor, & Power, 2002; Milanovic, Saville, & Shen, 1996). Investigations such as these, which shed light on the nature of rater variability, can directly inform rating procedures (e.g., rater training or revision of criteria), which, in turn, may ultimately enhance reliability and strengthen the weak link in an interpretive argument.

However, while investigation into assessor decision-making in evaluating tasks testing productive skills has flourished, to date there has been little research conducted on the decision-making behaviour of assessors in applying marking schemes to open-ended question types in the assessment of receptive skills. Although several recent research studies have explored the issues involved in designing automated scoring rubrics for open-ended items (see Carr, 2007; Carr, Pan, & Xi, 2002), many of the challenges discussed are unique to computer-based testing contexts. In order to address this particular gap in research on decision-making among human assessors on open-ended task types, a recent study by Harding and Ryan (2009) was conducted which examined markers' experiences while assessing Part B of the OET listening sub-test. The study identified three broad categories of decisions made during the assessment process:

1. *Decisions regarding spelling:*

   Referring to instances where an assessor must judge whether a misspelt response can still be accepted as evidence that the candidate has understood a particular piece of information in the text.

2. *Decisions regarding the correctness of an over-elaborate response:*

   Referring to instances in which candidates would write very long responses to short answer questions that included an unnecessary level of extra information (probably based on their general knowledge of the subject) within which assessors had to identify and evaluate a response.

3. *Decisions regarding the adequacy of response:*

   This was divided into two sub-categories:

   i. *Semantic distinction*
      Making a decision about whether a particular alternate word or phrase in a response demonstrated understanding of what the speaker had said, or whether it indicated a different concept which was not intended.

   ii. *Sufficiency of an answer*
      Making a decision about whether enough information was included in an alternate answer to match sufficiently the idea represented by the answer in the marking guide.

As well as identifying these types of decisions, Harding and Ryan (2009) also explored the bases on which assessors made decisions in these situations. It was demonstrated that assessors use a range of strategies to make their decisions which include utilizing resources

(such as dictionaries and medical reference texts), sharing knowledge and applying 'rules of thumb' (particularly in the case of spelling decisions). It was also shown that there was some degree of assessor variability, particularly in decisions concerning adequacy of response, when assessors diverged in their decision-making processes depending on their level of health literacy, which in turn influenced their perceptions of the underlying construct of the test.

2. Aim and research questions

The current study represents a replication of Harding and Ryan (2009), though with Part A of the OET listening sub-test as its focus. It is hypothesized that the nature of the note-taking task will result in some differences with respect to the nature of decision-making compared with that observed in the data concerning Part B. Specifically, the research questions we asked were:

1. What types of decisions are made by assessors while marking Part A of the OET listening sub-test?
2. On what bases do assessors make decisions?

3. Methods

To investigate these research questions, we conducted a stimulated recall group discussion (Gass & Mackey, 2000) with three OET assessors following their independent use of a marking guide in a routine administration of the OET. The technique allowed the assessors to identify where they had needed to make decisions not covered in the marking guide and to explain how they had done this. The group discussion format was seen as a natural choice in a context where raters routinely consult one another about marking dilemmas; however, it also provided an opportunity for assessors to disagree with one another about their marking decisions, and so display the bases of their decisions in a way which may not have been so easily accomplished through a series of individual stimulated recalls. Involving three participants provided a variety of comments that could be viewed as representative of the whole group of assessors without generating an overwhelming amount of data.

3.1 Participants

The three assessors who took part in the focus group discussion were Penny, Sarah, and Michael (pseudonyms). Penny also took part in the study reported in Harding and Ryan (2009); her professional experience provided an interesting counterpoint to the backgrounds of other assessors. While she also had a background in ESL literacy education, Penny differed from the other two assessors in that she was an experienced health science researcher. Sarah and Michael had both been teaching ESL for around ten years, and both had significant experience delivering exam preparation courses. All three were experienced assessors of English language examinations, though Sarah and Michael were relative novices in assessing the OET listening sub-test having only marked for one or two examinations at

the time the focus group was conducted. The discussion was moderated by two of the authors of the paper, who may be conceived of as participant observers, given that they are also developers of the OET listening sub-test. The OET assessment manager at the time (another author) was also peripherally involved in the discussion, but took on more of an observer's role. Note that these three authors were also the researchers who conducted the analysis of the data.

3.2 Materials

The test material which formed the basis for the discussion was the version used in a routine administration of the OET listening sub-test Part A. The version used in this study is known as 'Clare and the Optometrist', and the input text was a simulated consultation recorded between an actual optometrist, and a patient who had not been to the optometrist for two years. The test contained 62 items, divided under fourteen headings. Following administration (to 1126 candidates), the internal consistency of Part A was calculated as 0.88 (when combined with Part B, the reliability for the sub-test as a whole was 0.91).

3.3 Process of data collection

The procedure for collecting data replicated the method used in Harding and Ryan (2009). Three assessors were asked to volunteer to take part in the research project during a routine markers' meeting. When Sarah, Michael and Penny volunteered, they were invited to join in a focus group discussion after they had worked independently with the marking guide following the administration of a newly developed test. They were asked to keep any notes that they would routinely make on their copy of the marking guide, and to be mindful of any particular difficulties they had in applying the guide, or any points at which they felt they had to make a decision on a response that fell outside its parameters.

Following a routine marking period (approximately two weeks), the focus group was convened. Although this was a long interim between initial marking and verbal report, it would not have been possible to convene the focus group earlier without some impact on regular scoring procedures. For this reason the notes kept by assessors functioned both as a detailed record of decisions and as a useful mnemonic tool. The discussion lasted around two hours (including a break) and was video- and audio-recorded. The video-recording of the discussion was so that each speaker could be easily identified at the transcription stage.

At the beginning of the focus group interview, the procedure of the session was explained, followed by introductions. During this initial discussion, the assessors were asked three general questions:

1. Could you talk about your experience in ESL teaching and your experience as a marker of the OET?

2. What was your general experience marking Part A of the listening test this time?

3. Are there any particular issues you would like to raise about Part A of the listening test or the marking guide at this initial stage?

Following this introductory phase, the focus group discussion followed a structure in which we (the researchers/developers) worked through each item of the test in order, and asked the assessors to comment at any point where they recalled having difficulty. In this way, the marking guide itself provided a stimulus for recall of decision-making. This item-by-item stimulated recall was supplemented by the more detailed notes the assessors had each kept on their own sites of decision-making.

3.4 Data analysis

The focus group discussion was transcribed and analysed for content by the three researchers against the categories of decisions which had been identified in Harding and Ryan (2009). During this process, each researcher coded the transcript separately, and then a meeting was held to compare the results of independent coding. Two instances of differences in coding were resolved through discussion, and data which did not fit existing categories were accounted for through a collaborative revision which yielded a revised set of macro categories. These are discussed below, together with representative examples. Within each category, several 'bases for assessor decisions' were also identified (see research question 2). These are also discussed throughout the next section.

4. Findings

4.1 Overview

The focus group revealed that the assessors were generally satisfied with the marking guide and felt that in most cases it provided clear guidance for them to mark test papers with confidence. However, assessors did report 26 instances of decision-making across 24 of the 62 test items; that is, there were 26 instances where at least one assessor in the focus group recalled having to make a difficult decision about an item in one or more of the 100+ scripts he or she marked during the assessment period. Examples were found which fit each of the categories generated by the 2009 study. In addition, data were also found which required substantial modifications to the second of the existing categories and the addition of a sub-category under the third 'adequacy of response' category. Table 2 shows the list of revised categories, together with the frequency of particular decision types. The most common type of decision recalled was "sufficiency of an answer", followed by decisions on spelling:

[INSERT TABLE 2 ABOUT HERE]

In the sections below, each of the types of decision will be discussed in more detail, with illustrative examples provided of each (addressing Research Question 1). In addition, in examining these decision-types, the bases upon which assessors made decisions at these sites will be drawn out, with a view to understanding the range of influences on decision-making processes (addressing Research Question 2).

4.2 Spelling

Decisions regarding spelling were reported for some of the less common words heard in the text. The excerpt below illustrates the difficulty Michael faced in dealing with variations on the word 'weepy':

Excerpt 1 (FG: 338-353)

Michael: I recorded all these words: 'weeping' … 'weapy': w-e-a-p-y, okay … 'weeby': w-e-e-b-y ... 'webby', I accepted, 'weepy' with a double-e-p-y, I accepted … 'weeping', I accepted, and 'watering' I accepted. I did not accept ... oh, I accepted 'wippy', but I didn't accept 'wicky', 'weekly', or 'rippy'.

…

Researcher A: How do you apply a spelling rule … for for options like that?

Michael: Well, obviously if they've said, 'weeky' or 'weekly', they've … clearly they've misunderstood, because those are possible words ... or ... you know, quite plausible words in English … Similarly with 'rippy'. But ... if they've said 'watering', who knows if that's a listening problem or whether it's just a reasonable guess?

Michael appeared to be applying the 'rule of thumb' for spelling which had been agreed upon at previous marking meetings. This rule holds that a misspelling should be deemed acceptable provided it is a reasonably close phonemic match of the target word, and the meaning of the response remains clear. Thus 'wippy' would have been acceptable, but 'wicky' would not.

While the basis for decision-making on all spelling items appeared to be the rule of thumb, there was also evidence in the focus group data that assessors may draw on their knowledge of particular L1 phonologies in an attempt to understand the source of spelling mistakes. Excerpt 2 shows Penny talking on this point in reference to a candidate who wrote 'weeby' instead of 'weepy'.

Excerpt 2 (FG: 359-360)

Penny: No, the 'b' ... the 'b', it only comes in because of the Arabic. So ... they haven't misunderstood the word, they just don't know how to spell it.

However, application of this sort of knowledge is problematic in that assessors' knowledge of common perceptual errors of various L1 groups was necessarily limited by their own experience:

Excerpt 3 (FG: 383-383)

Researcher B: Does that, um … for other nationalities? Say, Korean – their 'f' and 'p' and all that sort of thing?

Penny: See I don't know a lot about that.

Michael: And, um ... it might also apply to Indian, because they have the three ... three sounds around the b-p area. In Indian ...

Penny: But, then again, if I've accepted it for some people, I'll accept it from others. You know, I won't *not* accept it.

Researcher A: Yeah, so they'll they'll just push out the limit a bit more, and then you'll ...

Penny: Yeah.

Irrespective of these different understandings of where perceptual errors may stem from, the three assessors appeared uniform in their application of the rule of thumb throughout the discussion.

The nature of the note-taking task allowing candidates to use their own words in responses means the issue of spelling variation is common (and often compounded by candidates' handwriting). The rule of thumb as applied in this context may not be ideal but it does provide an arbitrary standard of consistency among assessors, since assessors generally agree on how to apply it. Accommodation of misspelling based on individual assessors' knowledge of possible patterns of L1 interference is more problematic and is discussed in Section 5.

4.3 Supplementary information

In the current study a decision relating to an over-elaborate response was identified only once in the data; however, on discussing this segment the researchers felt it more appropriate to use the term 'acceptability' rather than 'correctness'. In addition, a related site of decision-making was identified which concerned the effect of an error in supplementary information on an otherwise acceptable response. In the current study, these were grouped together under a broad heading 'decisions regarding supplementary information'. Each sub-category is discussed in more detail below.

Acceptability of an over-elaborate response

Excerpt 4 below shows a clear example in which Michael recalled difficulty in evaluating the acceptability of an over-elaborate response. What is noteworthy about this example is that Michael's attitude towards such responses is quite unforgiving: once the response has been identified as 'from their own knowledge', the assessor is predisposed to being 'quite hard':

Excerpt 4 (FG: 592-603)

Michael: I had a problem with a few candidates – two or three – who gave quite detailed technical explanations of astigmatism, but didn't use the terminology of the speaker. I felt that they were giving answers from their own knowledge and not … so I tended to mark quite hard.

Researcher A: Um ... if the answers didn't bear any relation to what's ...

Michael: Well, they they were just, you know, rolling out an answer because somebody knew what astigmatism was and just rolled it out ... without using the same terminology ... I felt that …

Sarah: But if you change the word ...

Michael: I'm not talking about changing a word, I'm talking about people ...

Sarah: Okay

Michael: It was clear that they hadn't ... were not writing what they'd heard.

Michael appears certain that these responses are 'rolled out' based on their technicality, detail, and dissimilarity from the terminology used by the speaker. In other words, the incongruity of sophisticated language and complex thought within the context of other hastily written notes is a trigger for this type of decision. Yet, although it is the level and degree of supplementary information that is problematic, the decision ultimately concerns whether the *whole response* can be taken as evidence of the ability under test, as distinct from whether some parts of the detail is correct or not.

Effect of error in supplementary information

By contrast, excerpt 5 (below) shows a discussion in which the assessors consider the effect of erroneous information (according to the text) which is supplied adjacent to (and connected with) information that would ordinarily comprise a correct response according to the marking guide:

Excerpt 5 (FG: 407-426)

Michael:	There was another instance, just following on from this, with the weeping eyes ... and the candidate had written 'two days weepy eyes'. Now, in fact, the problem with weepy eyes had existed since last year. So the candidate's actually misunderstood what was said, however ... we have in previous cases discussed a policy – if they've given the right answer, it's in there, then that should be okay, providing they're not actually negating it. And, so I allowed that, even though the 'two days' bit was wrong.
Researcher A:	Mm
Researcher B:	Mm
Michael:	I'm not happy about doing it, but I did it ...
Researcher A:	[to Sarah] Would you have accepted that one?
Sarah:	Um ... probably, yeah. And you just discard …
Michael:	So I just did what I … feel the question ... I didn't think it was fair ... but I did it because that was …
Researcher A:	It wasn't a grave enough addition to ...
Michael:	To negate the meaning of the ...
Sarah:	'Cos you're right. They can write completely contradictory things within one line, so which ... at what point do you start saying, 'actually ...'

In this scenario, Michael appears to base his decision on the gravity of error in supplementary information – 'that should be okay, providing they're not actually negating it'. He also mentions some consideration of fairness to the candidate; although it is unclear whether he believes his decision was ultimately unfair, or whether he feels marking such a response as incorrect would be unfair.

The most difficult aspect of dealing with this type of response is that markers must have sufficient knowledge of the text in order to evaluate whether an error in supplementary information is serious enough to mean that correctly supplied information is no longer acceptable. For this reason, the familiarisation period in the markers' meeting

during which the assessors complete the test themselves is crucial; as is their ability to have recourse to the transcript. However, in a practical sense it would be difficult for assessors to make decisions such as these if they appeared often on what is usually a 60-70 item test. The focus group data suggest that the default position of the assessors was that unless there was a clear negation of the proposition, the answer would be accepted on the grounds that the candidate has heard the essential information (the information required according to the marking guide) correctly.

4.4 Adequacy of response

The same two types of adequacy of response decisions which emerged in our previous study – semantic distinction and sufficiency of an answer – were also found in the data for Part A. In addition, a third category emerged – acceptability of a blended response – which seemed to be unique to the nature of the note-taking task and is defined below. Consistent with the findings in Harding and Ryan (2009), decisions regarding adequacy of response appeared to be the most common source of divergence for markers. These types of decisions are often based on an individual assessor's understanding of the underlying construct of the test, and so, not surprisingly, discussion on these divergent points often revealed differences in assessors' own orientations towards the test. These are discussed in greater detail below.

Semantic distinction

Excerpt 6 shows a fairly clear decision over the semantic distinction between 'not soft' and 'rigid' contact lenses. It is noteworthy, however, because Penny disagrees with the synonymy of these two terms (with 'rigid' expressed as 'hard'):

Excerpt 6 (FG: 302-313)

Sarah:	I had someone put 'not soft lens'. I thought that was a good answer.
Researcher A:	Interesting. Did you accept that one or not?
Sarah:	I think I did, yeah.
Penny:	I didn't.
Researcher A:	Did you also have that?
Penny:	I remember having those ... I didn't accept it.
Researcher A:	Why, ah ... what ...
Penny:	Um ... 'not soft' doesn't necessarily mean 'hard'. It could be anything, I suppose. It just didn't ... we're looking for ... it's also a listening ... I mean, it is understanding. I mean, I think the understanding is in some ways more important than the listening, but, um ... well, they did understand, I suppose, I don't know. I just didn't accept it. I don't know. That was just my gut feeling.

In this example, Penny appears to vacillate in her rationale for not accepting 'not soft', but her comment 'we're looking for … it's also a listening … I mean, it is understanding' may point towards her greater orientation towards precision in candidates' responses as a key component of the ability being measured. In the data presented in Harding and Ryan (2009),

Penny discussed her health research background, which may have influenced the way she marks:

> … when I look at an answer that has that diverge from the marking guide, I try to work out whether the person understood what they were supposed to understand … I think my understanding about a lot of this is fairly good. (p. 111)

By contrast, Michael and Sarah attempt to understand responses such as these from the perspective of the candidate. In Michael's case, this was based on his own observation of 'normal' note-taking behaviour through his experience marking the test:

Excerpt 7 (FG: 319-320)

Michael: I found a number of ... a number of times where people were using the opposite ... they gave you the opposite, which was quite common.

In Sarah's case, the decision process was informed by her own experience as a language learner:

Excerpt 8 (FG: 321-323)

Sarah: 'Cos I think, sometimes I do a similar thing in note-taking ... if there's a word where I'm ... not the same in my own language, but it's a word I'm unsure of, I'll try to write it another way to try and get that mark.

Sufficiency of an answer

A similar pattern of divergence was found in an example of a decision concerning sufficiency of an answer. Excerpt 9 demonstrates again that Penny's particular background (in this case her knowledge of how medications can be delivered) informs her decision about whether the term 'drops' is an essential part of the concept that was expressed in the marking guide as 'antibiotic drops':

Excerpt 9 (FG: 1195-1215)

Penny: Oh, look, there was a real problem with the drops.

Researcher A: Yep, yep.

Penny: Because they'd say 'drops' and 'antibiotics' and I didn't accept it …

Michael: Mm, mm … very common.

Penny: … because it was supposed to be 'antibiotic drops'.

Researcher A: Mm … that's fair enough, I think. Yeah.

Michael: And a number of them just said 'anti … anti-inflammatories' and 'antibiotics', not mentioning the drops.

Penny: Yeah … yeah.

Michael: But I thought that was enough.

Penny: Oh, so you did. See, I didn't.

Michael: Well, 'cos it … that's what the answer is about, I felt.

Penny: No, because you can take it orally or you can put drops in your eye, and this is about an eye procedure, so it's about antibiotic drops not about antibiotics.

Sarah: I wasn't sure about that.

Penny: That's what I thought – that was my judgment.

Sarah: Yeah, that's how I've done it as well.

Researcher A: Mm.

Michael: I just felt the antibiotics, the 'AB' was the more important part of that, but maybe I'm wrong.

In this example it is evident that Penny and Michael are both trying to achieve the same goal – to assess the response based on its similarity to the underlying concept expressed by the marking guide entry. However, Penny is again oriented more towards precision on the basis of her professional knowledge. It should be noted that an exact reading of the marking guide makes 'antibiotic drops' the correct response. This is an instance where Michael makes his own decision although the marking guide in fact deals with exactly the point in question.

Acceptability of a blended response

The final sub-category – a decision type unique to the note-taking task format – was judging the acceptability of a 'blended response'. A blended response may be defined as separate pieces of information (i.e., concepts expressed as separate items in the marking guide) that are conflated, and perhaps condensed or recombined, in a candidate's response. Excerpt 10 shows a clear example of this phenomenon in the assessors' consideration of responses that blended two distinct items, 'cleans nightly' and 'stores in storage solution':

Excerpt 10 (FG: 463-472)

Michael: I found a lot of the answers blurred A, B, C ... got them mixed up.

Researcher A: OK.

Michael: I (inaudible) things like that.

Penny: Yeah, I ... there is some blurring that you can accept. I mean I'd accept something like 'stored nightly in solution'.

Researcher A: Mm hmm

Michael: Yep.

Penny: That would get ... ah ...

Michael: Two marks.

Penny: Probably two marks, yeah.

There is little controversy over this example and therefore no indication of how assessors made their decisions in such circumstances. However, it is worth noting that blended

responses occurred only for this particular set of items, suggesting that it was in fact the marking guide that was deficient in accurately representing the most common answers that assessors were likely to encounter. In these cases, the assessors are bringing to light a problem introduced in the development of the marking guide. Test developers might adjust the guide for use at subsequent test administrations; awareness of this issue might lead to similar situations being noticed and avoided during the development of marking guides for new tests.

4.5 Other influences on decision-making

Throughout the focus group, the assessors referred to a number of strategies they used when making decisions in addition to those which have been discussed above. These included talking to the assessment manager, talking to other assessors, and checking medical dictionaries (particularly for abbreviations). At other points, when the meaning of an answer was particularly unclear, the assessors seemed to orient towards a rigid application of the marking guide – even if this made them uncomfortable. Indeed, one of the striking characteristics of comments around the marking of notes is that, in judging numerous items of this kind, assessors feel a kind of sympathy with the candidate who gives an almost sufficient answer which is heightened by repeated experiences of clearly insufficient answers:

Excerpt 10 (FG: 632-633)

Sarah: But there were so many crappy answers that I just went, 'no way, no way', and then I went, 'ooh - like, that's so close, you know... I wanna give you that one.'

Although there was no evidence in comments such as these that assessors' affective response to the process of applying the marking guide had any tangible effect on their decisions, this is clearly a factor which should be considered as a potential influence on marking behaviour.

5. Summary and implications

This study has confirmed and expanded the findings of Harding and Ryan (2009) to develop a taxonomy of decision types which are made during the process of assessing Part A of the OET listening sub-test:

1. Decisions regarding spelling

2. Decisions regarding supplementary information

   i. Acceptability of an over-elaborate response
   ii. Effect of error in supplementary information

3. Decisions regarding adequacy of response

i. Semantic distinction
ii. Sufficiency of an answer
iii. Acceptability of a blended response

Revisions to the findings in Harding and Ryan (2009) involve changes in the structure and wording of category 2 and the addition of sub-categories 2.ii and 3.iii. These are probably due to the different nature of the Part A note-taking task: candidates have more opportunity to express their understanding of what they hear (or what they already know) in their own words without the physical constraints of a particular task as formatted on the page. There is similarly more scope for candidates to supply information that may contradict itself within the same response or that represents the meanings of the input material in different combinations from those specified in marking guide.

In addition to the decision types, several different rationales for, and influences on, these decisions have been identified in the data. These are summarised within three categories below:

Bases for decisions

- Application of rules:
    - general 'rule of thumb' for spelling
    - version-specific rules decided in marking meeting
- Utilization of resources:
    - discussion with other assessors
    - discussion with assessment manager
    - checking dictionaries, etc.
- Assessor knowledge and beliefs:
    - knowledge of the text (e.g., in determining gravity of error in supplementary information)
    - knowledge of subject matter and technical terminology (topic literacy and health literacy)
    - knowledge of language generally (e.g., in determining synonymy)
    - knowledge of patterns of L1 transfer (e.g., in evaluating misspelt words)
    - beliefs about the nature of listening and note-taking
    - beliefs about fairness
    - beliefs about the underlying construct of the test

This third label, 'assessor knowledge and beliefs', can be understood as similar in nature to Borg's (2003, p. 86) definition for 'teacher cognition', and incorporates any aspect of what individual assessors 'think, know or believe' in relation to their work.

The data show clearly that the individualised nature of assessors' bases for decisions may lead to variability in scoring a note-taking test, albeit in a relatively minimal way, given

the small number of instances reported. This, firstly, has obvious implications for assessor training. For example, the 'rule of thumb' approach to spelling decisions becomes problematic when individual assessors use their own knowledge of other languages and the effects of L1 transfer on English spelling. The inevitable variation in such knowledge across the group of assessors makes any standardisation impossible given the range of words (and of candidate language backgrounds) potentially in play for each version of the test. Such insight, not shared by all the assessors, is therefore not especially useful and its use may need to be limited. Secondly, and perhaps more importantly, the finding that assessors who are domain experts may mark more strictly than those who are not raises a broader question: who is the ideal assessor for a specific purpose language test? Rather than suggesting a particular assessor background as the 'ideal', we would argue that the aim should be to share knowledge on the topic of the test throughout the assessment process. It will never be possible to achieve homogeneity of experience and knowledge among any group of assessors; nevertheless, in a specific purpose test, assessors who are not subject experts may be encouraged towards a greater understanding of the topic. The presence of assessors with specialised knowledge should be viewed positively, provided that the relevant information is made available to all assessors in some way. Ideally, consistency will be achieved at a 'best possible' level for the group of assessors, which seeks an understanding of the topic similar to that expected of the candidates themselves.

Further research is required, therefore, to evaluate the usefulness of various strategies which help to ensure that assessors make the same decisions when they are faced with novel challenges. It would be reasonable to hypothesize that processes of sharing knowledge and discussion about beliefs are central to achieving a level of shared understanding. For these reasons, it would be useful to encourage greater levels of dialogue between test developers and assessors at the training stage in order to communicate the construct and rationale for decisions about what is in the marking guide more rigorously and more collaboratively, and also more interaction between assessors themselves at the early stages of the marking process regarding interpretation of the guide as well as particular knowledge of the topic of the test. In order to integrate this into a regular marking procedure, a 'stage one' check (as proposed in Harding and Ryan, 2009) could be implemented, taking the form of a second markers' meeting – after 10-20% of papers had been marked – during which initial decisions could be discussed and a joint decision made. This would be beneficial in terms of providing both a tangible outcome (a group decision) and a forum in which decision processes are articulated and scrutinised collaboratively.

Finally, as there is very little literature on the issue of assessor decision-making in marking notes (or open-ended items more broadly), this type of research should be conducted in other contexts, with different assessors and through different methods. The data presented in this study are highly situated – as a specific purpose language test with a very small team of assessors, the OET is not a case from which to make easy generalizations. However, the taxonomy of decisions would provide a useful departure point for further research seeking to investigate the bases of assessor decisions in a different context.

Acknowledgements

We would like to acknowledge the cooperation of the OET Centre, and in particular the assistance provided by Gerrard Neve. We would also like to thank the English Language Institute at the University of Michigan for providing financial support for this project through the Spaan Fellowship program.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Harding, L., & Ryan, K. (2009). Decision-making in marking open-ended listening test items: the case of the OET. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *7*, 99-114.

Bachman, L. F., Lynch, B., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12(2), 238-257.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Borg, S. (2003). Teacher cognition in language teaching: A review of research on what language teachers think, know, believe, and do. *Language Teaching*, 36, 81-109.

Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. *IELTS Research Reports*, 3, 49-85.

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

Carr, N. (2008). Decisions about automated scoring: What they mean for our constructs. In C. A. Chapelle, Y.-R. Chung, & J. Xu (eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 82-101). Ames, IA: Iowa State University.

Carr, N. T., Pan, M., & Xi, X. (2002, December). *Construct refinement and automated scoring in Web-based testing.* Symposium paper presented at the 24th Annual Language Testing Research Colloquium, Hong Kong.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, 86(i), 67-96.

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing*, 18(2), 133-147.

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge: Cambridge University Press.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155-185.

Elder, C., Harding, L., & Knoch, U. (2009). *OET reading revision study*. Final report. Language Testing Research Centre, University of Melbourne.

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

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

Kane, M. (2006). Validation. In R. L. Brennan (ed.), *Educational measurement* (4th ed.). Washington, DC: American Council on Education/Praeger.

Knoch, U. (2009, March). Investigating the effectiveness of individualized feedback for rating behaviour: A longitudinal study. Paper presented at the 31st Annual Language Testing Research Colloquium, Denver, Colorado.

Lumley, T., & Brown, A. (1996). Specific-purpose language performance tests: Task and interaction. *Australian Review of Applied Linguistics,* Series S(13), 105-136.

Lumley, T., Lynch, B., & McNamara, T. (1994). A new approach to standard-setting in language assessment. *Melbourne Papers in Language Testing*, 3(2), 19-40.

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications and training. *Language Testing*, 12(1), 54-71.

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

McNamara, T. F. (1990b). Item Response Theory and the validation of an ESP Test for Health Professionals. *Language Testing*, 7(1), 52-76.

McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing,* 8(2), 139-159.

McNamara, T. F. (1996). *Measuring second language performance*. London & New York: Addison Wesley Longman.

McNamara, T., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. Language Testing, 14(2), 140-156.

May, L. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing,* 11 (1)*,* 29-51.

Milanovic, M., Saville, N., & Shen, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Studies in Language Testing 3* (pp. 92-114). Cambridge: Cambridge University Press.

Orr, M. (2002). The FCE Speaking test: using rater reports to help interpret test scores. *System, 30*, 143-154.

Ryan, K. (2007). *Assessing the OET: The nurses' perspective*. Unpublished Masters thesis, The University of Melbourne.

Weir, C. J. (1993). *Understanding and developing language tests*. New York: Prentice Hall.

Xi, X. (2008). Methods of test validation. In E. Shohamy & N. H. Hornberger (eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., vol. 7, pp. 177-196). New York: Springer.

Table 1: Overview of the OET listening sub-test

	Input	Task(s)
Part A	Consultation between patient and health professional (approx. 25 minutes)	Note-taking under headings
Part B	Lecture/talk on health related topic (approx. 25 minutes)	Mixture of fixed-choice and open-ended items (including sentence completion, short answer questions, lecture note completion, chart/diagram completion)

Table 2: Frequency of decision types

Decision type	Frequency in data
1. Spelling	6
2. Supplementary information	2
(i) Acceptability of an over-elaborate response	(1)
(ii) Effect of error in supplementary information	(1)
3. Adequacy of response	18
(i) Semantic distinction	(3)
(ii) Sufficiency of an answer	(13)
(iii) Acceptability of a blended response	(2)

Figure 1: Example Part A task (with actual candidate's responses)

Question 3

Clare's contact lenses and details of the problem

- since 13 years old
- for the last year trouble
- scratchy, wippy,
- something rong with
- really just pain
-
-

Figure 2: Example of corresponding marking guidelines

Question 3 Clare's contact lenses and the details of the problem
1 mark for each of the following

3a	(wears) rigid (lenses)
3b	(worn) since 13 (years old)/for 15 years
3c	(feel) uncomfortable
3d	(a bit) scratchy (towards the end of the day)
3e	eyes weepy
3f	could be wrong care
3g	problems (only) recent/in last year