Exploring the validity of ESL reading test items by means of judgemental procedures

Susanna Ho Pui-san

INTRODUCTION

In Hong Kong, reading has always been an important component in two of the major public examinations: the Hong Kong Advanced Level (HKAL) Examination and the Hong Kong Certificate of Education (HKCE) Examination. Because of this emphasis in the two examinations, a considerable number of English teachers assign much of their class time to activities related to reading. With regard to reading comprehension and reading tests in particular, this paper attempts to answer the following questions:

1. How might reading comprehension be accurately measured?
2. What role does expert judgement play in content validation of a reading test?

To answer these questions, a reading test was designed and administered to 62 tertiary ESL students at the Hong Kong University of Science and Technology (HKUST). Nine teachers of English were asked to give their opinions of the test. The scores of the test takers and the views of the English teachers were analysed in detail.

MEASURING READING COMPREHENSION

A great deal of research has been done on reading tests, in particular on the types which measure test takers' cognitive abilities (Millman & Greene 1989:336). Cognitive skills, most of which are defined in various cognitive theories, form the basis of the testing points. Questions are set in such a way that test scores will reflect how well test takers have mastered the pre-selected cognitive skills. One kind of reading test that attempts to explore the readers' cognitive domain assumes, or attempts to prove, the existence of reading subskills. Early studies, such as that conducted by Davis (1968), demonstrated the significant role played by reading subskills in comprehension. Davis' study was conducted on a large scale (320 test items and 988 subjects). He found the existence of eight subskills which were 'independent of one another' (ibid.:542). The theories of reading subskills put forward in those early studies gained widespread recognition in the 1970s. Reading
tests were designed in accordance with the reading subskill framework; test items were divided into clusters, each of which tested a reading subskill. This kind of test had a great influence on syllabus design. To ensure that English teachers who were preparing students for reading examinations and/or tests were assisting them to acquire the necessary reading subskills, these were listed in syllabuses for teacher guidance.

The notion of reading subskills was brought into doubt when later research questioned their significance. Studies such as that conducted by Lunzer et al. (1979) refuted the concept of reading comprehension being ‘manifold’ and argued instead that it was ‘unitary’. They chose eight subskills (different from those listed by Davis) for their test, which was conducted on 257 subjects. Four different texts were used to test the eight different pre-selected reading subskills. These eight subskills were further divided into two types: those that required lower-order processing skills, such as the subskill of recognising word meaning, and those that required higher-order processing skills, such as inferencing. Factor analysis of the test scores showed that reading comprehension was a unitary concept.

A more recent study by Rost (1993) on first language reading comprehension also sheds some light on this area. Although he does not find ‘multiple dimensions in L1 reading comprehension’ (ibid.:89), he concludes that successful reading may be attributed to the ability to make inferences and interpret vocabulary correctly in a reading text.

**Characteristics of a good reading test**

One of the most important test specifications is to decide on the test format which best suits the test purpose. Allan (1992:101) explores the issue of testwiseness — which is defined as ‘the ability to use test-taking strategies to select the correct response in multiple-choice tests without necessarily knowing the content or using the skill that is being tested’ — in EFL/ESL reading test takers. In multiple-choice questions, test takers may derive hints from distractors in the question or they may make use of their knowledge of grammar to answer the question or they may just choose the ‘odd man out’, that is, an option which looks different from the others in terms of length, tense, etc. In order to avoid such possible problems caused by multiple-choice questions, free-response questions may be a better alternative.

Expert advice has always been regarded as important in test development. Its main function is to arrive at the ‘external validation’ of a test (Millman & Greene 1989:346). Objective views on a test are invited from teachers and/or experts in the field at different stages of test development. Their function is to improve the quality
of the test and thus increase its credibility, but not to make drastic changes in it. Different results in terms of the level of agreement among panels of judges have emerged from different studies. For instance, Alderson & Lukmani’s nine judges (1989) largely disagreed with each other, whereas Lumley’s five judges (1993) came to almost total consensus. In view of these diverse results, it was worth seeking expert advice for the present small-scale study.

**METHODOLOGY**

This study attempts to gather two major types of findings: test scores from the reading test and opinions from expert teachers. The ultimate purpose of the test was to examine how well first-year undergraduates at HKUST read. A secondary purpose of the test was to investigate whether reading subskills exist, that is, whether reading comprehension is unitary or manifold (Lunzer et al. 1979:37).

**Subjects**

The test used in this study was a proficiency test given in Fall 1993, to 62 first-year undergraduates studying at HKUST, all of whom were taking the English Language Enhancement Course offered by the Language Centre of the University at the time the study was carried out. The students of this course had obtained grade D or below in the Use of English paper of the HKAL examination. To make results comparable, the subjects for the two pilot tests were also selected from the same population.

**Procedure**

**Test construction**

Based on the different lists of reading subskills from previous studies (Davis 1968; Lunzer et al. 1979; Alderson & Lukmani 1989; Lumley 1993), a list of four reading subskills was established:

1. finding explicitly stated answers to questions;
2. deducing meaning implied by the author;
3. interpreting metaphors;
4. finding salient or main points.
Among the four subskills, finding explicitly stated answers to questions is believed to be the least cognitively demanding as readers are able to locate an answer directly from the text.

In this study, subjects read a text entitled ‘Nam’s New Wave’, which was taken from a local newspaper (Appendix 1). After reading the text, they were required to answer a set of free-response comprehension questions. In order to make the most of the free-response question type, spelling mistakes and grammatical errors, such as wrong tenses, made in the answers were not penalised, so that test takers with poor writing abilities were not discriminated against. Test items were written with the chosen reading subskills in mind. The first version of the test, which contained 32 items, eight items for each subskill, was piloted on a group of 24 students who had similar English proficiency level to those who would take the final version of the test. Among the 32 items, four seemed problematic as no, or very few, students got them correct. Two of these four items were rewritten and re-tested on another group of nine students. As the second group was given only fifteen minutes to finish the questions, results were very poor. Only one student answered one of the items correctly. In order to explore the reasons why these two items were so much more difficult than the others, they were kept in the final version of the test to be tried out on a bigger group.

Test administration

The final version of the test contained 30 items (eight of them testing literal comprehension, eight testing making inferences, seven testing the interpretation of metaphors and seven testing the identification of main points). The test was administered to 62 first-year undergraduates, all except one of whom were able to finish the test within 75 minutes. The general response from the test takers suggested that the text was difficult to understand and that there were too many test items.

The test papers were collected from the test takers and marked by the same marker, who followed a set of model answers. The answers were prepared on the basis of the setter’s own judgement and the alternatives given by the test takers in the pilot tests. To ensure fairness, all items carried equal weighting so that no one particular subskill was favoured. The test was also given to a group of nine judges to collect their opinions on the test design.

Judgemental phase

A three-stage judgemental phase was conducted in the hope of collecting expert opinions on the test in particular and reading comprehension in general. During the first stage, nine judges were asked to evaluate each of the questions in the final
version of the reading comprehension test and state which reading subskill(s) each of them was testing. The nine judges were Language Instructors working at the Language Centre of HKUST. All of them had had considerable experience in teaching ESL. Five were native speakers of English while four were Cantonese speakers. A set of materials, including the text itself, 30 comprehension questions, the answers to the questions and a table which judges were asked to fill in to indicate their choice of subskill, was given to each of the nine judges. They were told that the answers were for their reference only.

During the second stage, which took place after an interval of ten days following the initial stage, the same judges were asked to evaluate the same text and indicate which subskill(s) each question was testing. This stage was similar to the initial stage except that the judges were asked to put down their choices on a different table. This time, they were also asked to rate the level of difficulty of each question from easy to difficult and they were given a choice of four reading subskills instead of six. The two subskills omitted from the list were not tested in this study. They were on the initial list so as to give the judges more choices. However, the results generated from the initial stage showed that a good deal of the disagreement among the judges centred on these two subskills. In order to allow the judges to be more focused on the reading subskills being tested, these two were taken out in the second stage.

The third stage, which was an interview, took place a few days after the judges had finished the second judgement. Each of the nine judges was interviewed for fifteen minutes so that they could give their opinions of the test, to see if they made the same choices that they had made in the second stage. This stage proved essential as three of the judges did change their minds and chose different subskills for some of the test items during the interview. In the audio-recorded interviews, the nine judges were asked the same questions. They were asked to explain why they had chosen the subskill(s) for six of the 30 questions. The interviews were based on what the judges had chosen during the second stage. The six questions were selected for discussion for two reasons: either there was a lot of disagreement among the judges themselves in the initial stage, or the judges' choice of subskills was very different from the subskill intended to be tested. In the interviews, they were also asked to comment on the choice of text, and the design of the test, which included both the questions and the answers. Finally, they were also asked whether they thought that the mastery of reading subskills equalled the mastery of reading comprehension.
FINDINGS AND ANALYSIS

Judges’ opinions on the test

Which reading subskill is this item testing?

Six items were selected for discussion. Table 1 lists the judges’ choices in the initial and the second stages, and in the interview if they changed their minds.

Table 1: Judges’ choice of subskills for questions discussed in interviews.

<table>
<thead>
<tr>
<th>Q</th>
<th>J1</th>
<th>J2</th>
<th>J3</th>
<th>J4</th>
<th>J5</th>
<th>J6</th>
<th>J7</th>
<th>J8</th>
<th>J9</th>
<th>Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mi</td>
<td>I</td>
<td>EE</td>
<td>Mi</td>
<td>EE</td>
<td>EE</td>
<td>EE</td>
<td>EE</td>
<td>E I</td>
<td>Mi</td>
</tr>
<tr>
<td>3</td>
<td>Mi</td>
<td>Mi</td>
<td>W/Me</td>
<td>Me</td>
<td>Me</td>
<td>EE</td>
<td>E Me</td>
<td>Me</td>
<td>E Me</td>
<td>E Me</td>
</tr>
<tr>
<td>7</td>
<td>J I</td>
<td>EE</td>
<td>EE</td>
<td>EE</td>
<td>I E</td>
<td>EE</td>
<td>J I</td>
<td>I E</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>10</td>
<td>EE</td>
<td>EE</td>
<td>EE</td>
<td>EE</td>
<td>EE</td>
<td>EE</td>
<td>EE</td>
<td>Me</td>
<td>E E</td>
<td>I</td>
</tr>
<tr>
<td>18</td>
<td>EE</td>
<td>Mi</td>
<td>EE</td>
<td>EE</td>
<td>E I</td>
<td>E I</td>
<td>E I</td>
<td>J I</td>
<td>I I</td>
<td>I I</td>
</tr>
<tr>
<td>25</td>
<td>J I</td>
<td>Me</td>
<td>Me</td>
<td>Me</td>
<td>Me</td>
<td>Me</td>
<td>Me</td>
<td>Me</td>
<td>I /Me</td>
<td>Me</td>
</tr>
<tr>
<td></td>
<td>Me</td>
<td>Me</td>
<td>Me</td>
<td>Me</td>
<td>Me</td>
<td>Me</td>
<td>Mi</td>
<td>Mi</td>
<td>Mi</td>
<td>Mi</td>
</tr>
</tbody>
</table>

Key: Q - Question number
     Mi - Finding salient or main points
     E - Literal comprehension
     I - Drawing inferences
     Me - Interpretation of metaphors
     J - Forming judgements
     W - Word meaning
     Int. - Subskills intended to be tested
     J1–J9 - Judges 1 to 9

From the table, we can see that the judges were not completely consistent in terms of choosing the same subskill during the first and the second stages. This may suggest that this kind of decision-making is somewhat arbitrary. For items 2 and 25, no one chose the subskills that were meant to be tested in the initial stage. But in the second stage, some of the judges changed their minds and indicated that these two items were testing the target subskills. Interestingly enough, none of the judges indicated
that item 10 was testing the intended subskill, which was that of drawing inferences and all except one said that the question was testing literal comprehension:

**Item 10**

Why did the four Vietnamese take part in the Saigon Floating Hotel Surf Pro?

Section containing the answer: Admittedly, the four Vietnamese taking part in the Saigon Floating Hotel Pro, who were there courtesy of a wild card on the grounds that it was their beach, didn’t pull off the manoeuvres as well as the 32 other professionals.

Some of the judges pointed out that it was too lenient to accept the phrase *it was their beach* as the answer to the question. This kind of direct copying requires no particular skill except literal comprehension. Two of the judges mentioned that the phrase *courtesy of a wild card* was crucial to arriving at the answer, so one of them indicated that this question was testing interpretation of metaphor. The other judge felt that the question was testing literal comprehension but said that if test takers did not understand this phrase, they would come to ‘a mental block’ in comprehending the remaining part of the sentence. As the judges placed different emphases on different clues, this may explain why they could not come to a complete consensus as to which subskill each of the questions was testing.

Another reason why the judges could not come to a consensus is perhaps due to their own interpretation of what the subskills are. Although a short description of each subskill was given, some of the judges were still preoccupied with their own definitions of subskills, which could differ enormously from those of the other judges, depending on individual experience. One of them interpreted literal comprehension as test takers’ copying or borrowing words or phrases from the text. When test takers could not use words from the text in their answers, she would interpret this as making inferences. This unique interpretation, which is somewhat different from the original description of the subskill, may have led this judge to choose a different subskill from the other judges.

**Difficulty of test items**

The judges were also asked to state the level of difficulty of each item, as well as classifying it in terms of the subskill it was testing. Although test scores revealed that the test was difficult for the test takers, the judges themselves could not come to total agreement on this point. The following table shows the average ratings of each of the judges.
Table 2: Judges' estimations of test difficulty.

<table>
<thead>
<tr>
<th>Judge</th>
<th>Average rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.9</td>
</tr>
<tr>
<td>2</td>
<td>3.9</td>
</tr>
<tr>
<td>3</td>
<td>2.8</td>
</tr>
<tr>
<td>4</td>
<td>3.6</td>
</tr>
<tr>
<td>5</td>
<td>2.9</td>
</tr>
<tr>
<td>6</td>
<td>2.0</td>
</tr>
<tr>
<td>7</td>
<td>3.1</td>
</tr>
<tr>
<td>8</td>
<td>3.5</td>
</tr>
<tr>
<td>9</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Key: 0.5 – 1.4 – Easy
     1.5 – 2.4 – Fairly easy
     2.5 – 3.4 – Average
     3.5 – 4.4 – Fairly difficult
     4.5 – 5.0 – Difficult

We can see that four of the judges rated the test as 'fairly difficult', three rated it as 'average' and two of them rated it as 'fairly easy'. As the judges' ratings of the difficulty level varied, it may be more meaningful to look at the ratings of each of the test items.

When looking at the individual test items, seven of them were rated as 'fairly difficult' and two of them as 'difficult'. Among these nine items, only four of them were indicated as 'difficult' by both the judges and the test scores. They are items 1, 11, 14 and 23, three of which were testing the subskill of interpretation of metaphor and the other the subskill of making inferences. General consensus in terms of the difficulty level of these four test items was reached among the nine judges. Table 3 gives a summary of the judges' ratings of these four test items.
Table 3: Judges' ratings of item difficulty.

<table>
<thead>
<tr>
<th>Item/Judge</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>E</td>
<td>E</td>
<td>D</td>
<td>E</td>
<td>B</td>
<td>Me</td>
</tr>
<tr>
<td>11</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>C</td>
<td>E</td>
<td>E</td>
<td>C</td>
<td>Me</td>
</tr>
<tr>
<td>14</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>B</td>
<td>D</td>
<td>B</td>
<td>C</td>
<td>Me</td>
</tr>
<tr>
<td>23</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>C</td>
<td>E</td>
<td>E</td>
<td>C</td>
<td>I</td>
</tr>
</tbody>
</table>

Key:  
B  – Fairly easy  
C  – Average  
D  – Fairly difficult  
E  – Difficult  
Me – Interpretation of metaphor  
I  – Making inferences  
Int. – Subskills intended to be tested

Both items 11 and 23 were rated as 'difficult' by all judges except two. As these two judges were the only ones who rated the whole test as 'fairly easy' (Table 2), we might interpret this as meaning that their average items would be rated as 'difficult' by other judges. As for item 1, most judges rated it either 'fairly difficult' or 'difficult', with one of the judges rating it 'average' and another 'fairly easy'. Again, the one who rated this item 'fairly easy' rated the whole text 'fairly easy'. Most of the judges (five out of nine) rated item 14 as 'difficult', one rated it 'fairly difficult', one rated it 'average' and two others rated it 'fairly easy'. An interesting observation is that seven out of eight questions which were supposed to be testing the subskill of making inferences were interpreted by some of the judges as testing literal comprehension. As discussed earlier, test item 10 requires a straightforward answer; and thus fails to test test takers' ability to draw inferences from what they read in the text. It could be concluded that the level of difficulty of an item largely depends on the answer expected from the test takers. Further comments given by one of the judges reflect that a literal comprehension question could be made difficult by the wording of the question and the order of the question among all the other questions.

Consistency of judgements

The nine judges were asked to state which reading subskill each of the items was testing at two different stages (with a lapse of ten days). The judges were given a choice of six subskills (two of which are not included in the test) the first time, and a choice of four (only the subskills used in the test) the second time. Because of this difference, intra-rater correlation cannot be calculated in a very scientific manner.
However, a comparison of the judges’ subskill choices still reveals that some of them are more consistent than the others. The number of items that were classified as testing the same subskill during the first and the second time ranges from 13 to 22. In other words, the most consistent judge has the same opinion in terms of the subskill for 73% of the test items in both the first and the second stages. The average number of items for which the nine judges make the same choice in both stages is 15.8 (out of 30 items). The numbers show that consistency in choosing the same subskill for the same test item varies among individual judges.

*Reading comprehension: The judges’ point of view*

When asked whether the mastery of reading subskills equals the mastery of reading comprehension, four of the nine judges responded negatively, two positively, two responded positively, though with some reservations, and one claimed not to know enough about the subject to answer the question. Those who responded negatively preferred global understanding of the text to breaking down the text into discrete units. For teaching and testing purposes, using subskills as categories for measuring performance is still convenient. Two of the judges stated that the mastery of reading subskills reflected the mastery of reading comprehension to a certain extent. Those who showed a mastery of reading subskills were obviously better readers, although different text types might require different reading subskills.

The expert advice on the reading test serves as a preparation for analysing the test scores. While quantitative data give us the whole picture of the test in a uniform manner, unique responses from the judges (see Appendix 2 for a sample of interview extracts) explain and supplement the statistical data.

*Test results*

*Overall test performance*

The pilot test, involving 24 first-year undergraduates, yielded quite poor test performances, with an overall average of 33%. After revision, the final version of the test was administered to 62 first-year undergraduates. The scores were recorded and computed through the program, *Itemdemo*. This classical item-analysis approach gives a detailed breakdown of item difficulty and discrimination.

Table 4 gives a summary of test takers’ performance on the 15 problematic items. These items were either too easy or too difficult, or they were not able to discriminate between the capable test takers and the poor ones.
Table 4: Test results.

<table>
<thead>
<tr>
<th>Item</th>
<th>Subskill to be tested</th>
<th>Facility</th>
<th>Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Metaphor</td>
<td>0.226</td>
<td>0.471</td>
</tr>
<tr>
<td>2</td>
<td>Inference</td>
<td>0.129</td>
<td>0.235</td>
</tr>
<tr>
<td>4</td>
<td>Literal comprehension</td>
<td>0.226</td>
<td>0.059</td>
</tr>
<tr>
<td>5</td>
<td>Literal comprehension</td>
<td>0.242</td>
<td>0.176</td>
</tr>
<tr>
<td>6</td>
<td>Metaphor</td>
<td>0.758</td>
<td>0.235</td>
</tr>
<tr>
<td>11</td>
<td>Metaphor</td>
<td>0.016</td>
<td>0.000</td>
</tr>
<tr>
<td>14</td>
<td>Metaphor</td>
<td>0.113</td>
<td>0.235</td>
</tr>
<tr>
<td>15</td>
<td>Literal comprehension</td>
<td>0.677</td>
<td>0.176</td>
</tr>
<tr>
<td>16</td>
<td>Inference</td>
<td>0.323</td>
<td>0.059</td>
</tr>
<tr>
<td>18</td>
<td>Inference</td>
<td>0.161</td>
<td>0.294</td>
</tr>
<tr>
<td>21</td>
<td>Inference</td>
<td>0.758</td>
<td>0.294</td>
</tr>
<tr>
<td>22</td>
<td>Literal comprehension</td>
<td>0.161</td>
<td>0.294</td>
</tr>
<tr>
<td>23</td>
<td>Inference</td>
<td>0.081</td>
<td>0.059</td>
</tr>
<tr>
<td>26</td>
<td>Main point</td>
<td>0.145</td>
<td>0.176</td>
</tr>
<tr>
<td>29</td>
<td>Main point</td>
<td>0.177</td>
<td>0.529</td>
</tr>
</tbody>
</table>

Average of totals : 10.84
Median of totals : 10.00
Standard of deviation of totals : 4.01
Lowest total score : 3.00
Highest total score : 21.00
Kuder-Richardson (20) : 0.66
Number of participants in test : 62

Overall, the test is quite difficult, as we can see from both the test scores and the judges' ratings of item difficulty. Table 4 shows that 11 out of 30 items have a facility value lower than 0.25; this means that fewer than 25% of the test takers answered these items correctly. Among these eleven items, three test the ability to interpret metaphors, three test the skill of making inferences, three test literal comprehension and two test the ability to find main points. As far as test scores are concerned, it seems that no particular subskill is favoured, that is, no one particular subskill seems to be comparatively more or less difficult to master than any other. However, when the judges' opinions are also taken into account, interpretation of metaphor seems to be a more difficult subskill than the others.

From Table 4, we can see that the scores range from 3 (10%) to 21 (70%) with a median of 10 (33.3%). Although the test is difficult, it is quite reliable (the Kuder-
Richardson value being 0.66). The test seems to be measuring reading comprehension in terms of test takers' mastery level of the four reading subskills.

Some problematic items

On the whole, the test, except for items 15 and 16, helped to discriminate between the test takers' reading abilities. Item 15 was originally designed to test students' literal comprehension. Some interesting findings were obtained in the judgemental phase. During the first phase, eight out of nine judges indicated that the item was testing literal comprehension; however, when they read the text and the item for the second time, four of them changed their minds, with three switching to making inferences and one to finding the main point. As some of the judges changed their minds when they read the item for the second time, the item might not be as easy as had been thought, and thus it is not surprising that it 'tricked' some of the more capable students.

Item 16 was designed to test students' ability in making inferences. During the first judgemental phase, only three judges chose this skill and among them, two were undecided: one of them could not decide whether the item tested this skill or literal comprehension, whereas the other could not decide whether it was testing making inferences or word meaning. Again, this question 'tricked' both the top and the average test takers.

Among all the thirty items, item 11 was the most difficult. This item, which asked test takers to explain the phrase 'courtesy of a wild card' in their own words, was first designed to test students' ability to interpret a metaphor. Seven judges agreed that this item was testing the intended subskill. As for level of difficulty, seven of them rated the item as difficult and the other two rated it as average. A tentative conclusion could be drawn that this item was seen as difficult by both the judges and the test takers (through their test performance). As only one test taker got this item correct, its discriminatory power is insignificant. In fact, the item is not discriminative at all as the test taker who got it correct scored a below-average mark of only 20%.
DISCUSSION

Accounting for text difficulty

This text is very culture-specific, with several examples of American slang. Although it is true that the test takers are very likely to come across this kind of text in real life, it is doubtful whether they are really interested in reading a text like this, which requires a good deal of background knowledge. To keep test takers interested in the text, and hence in the test, one of the judges suggested choosing a text topic with a local flavour. In this way, test takers will have a better chance of comprehending the text successfully. While topic is an issue in reading test construction, it is worth mentioning that in Hong Kong’s public examinations, authentic reading materials on various topics, both local and foreign, are used. These reading materials are adapted and modified to match the level of difficulty with the candidates’ proficiency.

Which subskill is comparatively more difficult?

From both the test scores and the judges’ opinions, it was found that among the four most difficult items, three of them were testing the test takers’ ability to interpret a metaphor. The fact that this subskill is shown to be comparatively more difficult than the other three can be attributed to two possible causes. The ability to interpret a metaphor and appreciate the language takes time. As the test takers do not have enough time to read the text over and over again, they may not be able to appreciate the language. Second, to understand a metaphor is a risk-taking task. To understand a metaphor is to understand a word or phrase which is not expressed in the everyday sense. In daily reading, we are more willing to take risks. However, in a test, test takers are less ready to take risks because the wrong interpretation of a metaphor will lower the test score.

Test item design

This study shows that literal comprehension, which is considered to be a comparatively less demanding reading subskill, might still pose some difficulty to certain test takers, depending on such factors as how the questions are phrased. A test item could be made more or less difficult by the wording of the question and the sequence of the item as a whole. The questions should be clear and concise, without creating ambiguity for the test takers. In this study, the judges helped to identify the ambiguities of the questions if there were any. The 30 test items are all free-response
questions. With no other test format, the test does not reflect reality. Most of the public examinations found in Hong Kong contain test items in various formats, for example, summary writing, multiple-choice questions as well as free-response questions. With different test formats, the test can meet the different preferences of the test takers. It always seems that some test takers prefer one type of test format to the others. With only one type of test format, some test takers, especially those with poor writing abilities, may be discriminated against.

In the present test, the thirty questions are sequenced in line with the text order. In other words, the test takers do not have to read back and forth in the text to look for the answers. Despite the convenience of this arrangement for the test takers, there are two drawbacks to this sequencing. Once the test takers discover this sequencing, they can make use of their test-wiseness and get the correct answers without too much detailed reading. This may defeat the purpose of testing students’ reading comprehension. Another drawback of this kind of sequencing is that the items are not presented to the test takers in order of increasing difficulty. When the first item in a test is difficult, which is the case here, the test takers may feel frustrated and may not be willing to continue.

*Judges’ opinions on the test*

As an essential part of the study, the judgemental phase yields valuable opinions on the test and on theories of reading comprehension. The nine judges gave quite comprehensive and detailed comments on the test. However, since they are all teaching in the same tertiary institution, they do not represent a cross-section of English teachers in Hong Kong. Because of the tight schedule of the Language Enhancement Course run by the Language Centre at HKUST, the final test and the judgemental phase were conducted simultaneously, which meant that even if problems had been spotted by the judges, nothing could be done to improve or change the test content. It was found that some of the judges were dissatisfied with the answer to item 7. It was pointed out that the text did not provide sufficient information to answer this question. It would certainly have been preferable to have conducted the judgemental phase prior to the final test.

Some test items were seen as testing *literal comprehension* by some of the judges and as testing *making inferences* by the others. These diverse results show that they did not agree that all of the items have the same level of difficulty. To have a better understanding of these items, they were chosen for discussion in the interviews. It was found that the judges had different expectations from the test takers. The judges who assumed the test takers had a wide range of vocabulary and sufficient background knowledge, would identify a certain item as testing *literal comprehension*. 
Those who had low expectations of the test takers would choose *making inferences* for the same item. This information, gathered in the interview, explains why certain items were viewed so differently in terms of the difficulty level by the different judges.

It was also found that all but two of the judges rated the test as 'fairly difficult'. As there were individual differences among the judges, when we look at the ratings given by each of the judges on item difficulty level, we need to consider their expectations of the test takers. Obviously, two of the judges in the study had high expectations of the students and thought that they would be able to answer most of the questions. In view of such individual differences, the need for interviewing the judges arises. The different opinions given by the different judges in the interview reflect reality. In the real world, each English teacher may bring his/her own beliefs into teaching methodology and classroom practice and this may result in different emphases being put on the different aspects of a reading classroom. While it is true that in reality, teachers bring their different opinions into the reading classroom, it would be difficult for them to standardise the specifications of a test if they insisted on their own beliefs. This study, which involves nine judges, seems to have failed to foresee a large group of teachers having such diverse opinions.

With such diverse opinions on the subskills, we cannot conclude which subskill each item is testing. What we can do with the test scores is distinguish the more competent readers from the less competent ones. Perhaps, the extremely low test scores should be eliminated from the data profile as the test takers concerned may have no interest in doing the test and their test scores may have distorted the implications presented by the data. Alternatively, we may interview these low scorers to gain more information about the reasons for their poor test performance.

**CONCLUSION**

To conclude, let us review the two questions that were stated in the Introduction:

1. How might reading comprehension be accurately measured?
2. What role does expert judgement play in content validation of a reading test?

Now, let us examine how far the findings derived from this study are able to illuminate these questions. This study seems to support the claims of Lunzer et al. (1979) that reading comprehension is not a multiple of reading subskills; instead, it is a cognitive activity that is unique to every individual. Depending on a reader's past experience and world knowledge, the same test could demand different skills from
different readers. As mentioned in the last section, we can see that the difficulty level of a test item depends on how demanding the answers are. Therefore, we cannot conclude that some reading subskills are more difficult to master than others in a reading test; we have to realise instead, that the level of difficulty varies from item to item. From this study, we can see that reading comprehension can be measured by means of a reading test with various formats of test items. However, even when every test specification is carefully attended to, a test may still create problems if the reading text covers an alien topic, or the test contains too many items. A more valid test is thus made up of a reading text with a topic that is familiar to target test takers and a reasonable number of items of various types. As for the second question, we can see that judgemental validation is essential to making sure that the test items match the test specifications, and that the answers to the items are correct.

In this study, the test takers performed poorly. This could be due to two main reasons. One is that the test takers were not proficient in English as they scored grade D or below in the HKAL examination. Another reason may be that they did not have sufficient prior knowledge of Vietnam to activate the necessary schemata to work out the test. To overcome this difficulty, teachers should provide or activate the necessary background knowledge to help students to understand a text in a reading classroom.

This study seems to confirm the important place held by a group of expert teachers who give external validation of the reading test items. In some cases however, we may need very few judges for test validation (cf. Millman & Greene 1989). Also, it is not necessary for all the judges to come to a consensus on all matters in test construction. In future research, it might be better to invite a smaller group of teachers and/or language experts to comment on the test. In the real world, three to five moderators from various domains such as secondary schools and universities are invited to comment on the examination drafts for public examinations. This practice is to ensure a fair and objective test. What we may also do in future is consider comments from test takers, too. We can ask tertiary learners for their opinions on the test after they have finished it. By interviewing them, we could be made more aware of the factors that make a test difficult for test takers. One final word on the study is that since it fails to support the existence of reading subskills, and in view of the trend of integrating different skills (reading, writing and even listening) within the same examination paper, it would be sensible to design a test that integrates two or three skills. To make a test valid and reliable, its test method and instruments, markers and moderators, test subjects and data analysis method are important factors to consider during the course of test construction.
Appendix 1

NAM'S NEW WAVE

Immortalised in countless Vietnam War films, China Beach has at last played host to those who most worship it — the surfers.

1 'CHARLIE don't surf,' says the Robert Duvall character of the Vietnamese in Apocalypse Now. Well, now he does. Last weekend, in scenes as reminiscently cliched as the whole '60s 'thing', long-haired California types descended on a stretch of white sand known as China Beach and shared a few waves with the locals.

2 The circumstances were as extraordinary as the setting was incongruous; this once ravaged country and its people playing host to the very participants, and their descendants, who did the ravaging. This time the foreign invaders came armed not with bombs, but surf boards and peace symbols.

3 For Hong Kong resident Dave Garcia, who hatched the idea of the surf competition with friend Bruce Atten of Sport Asia, it was like unleashing a tidal wave of memories of his youth, which had long been held at bay.

4 He had surfed China Beach, the infamous rest and relaxation resort, during the war. As a GI he took part in the assault on nearby Hamburger Hill, where a large number of American soldiers were slaughtered in the attack, and made an emotional pilgrimage back there during the event.

5 'It brought back some heavy memories, but life must go on,' he said. 'We found what we thought was the hill, but we couldn't find the exact spot where we were because it had changed so much. It used to be thick jungle and now it's just tiered forest. We had also been on a 17-hour drinking binge and so we had changed quite a lot as well'.

6 'The nice part was to go back and do the surfing. I had some real fond memories of relaxing at China Beach and it was great to get back there'.

7 Surfing had hardly been a cultural exchange during the war. During the past 20 years few boards have been spotted, bar the occasional visitor in the recent wave of tourism and a bunch of American high school scholars who brought a few along on a field trip last December. But Atten and Garcia had successfully helped stage the Ho Chi Minh and Hanoi marathons, so why not surfing?

8 One year later, dozens of beach types frolicked in the murky waters off Da Nang (the brochures described it as crystal clear and deep blue — beyond even the exaggerations of an estate agent), and performed such surfer-speak tricks as 'cut-backs', 'bottom turns', 'floaters' and 're-entries'. Admittedly, the four Vietnamese taking part in the Saigon Floating Hotel Surf Pro, who were there courtesy of a wild card on the grounds that it was their beach, didn't pull off the manoeuvres as well as the 32 other professionals.
9 Their competition came to Da Nang from Brazil, Australia, Tahiti, Japan and, of course, California. So seriously is the sport taken that pro-circuit participants are prepared to beg, steal and borrow to pay for air fares, hotels and freight expenses to move themselves and up to six boards, from one beach to another around the world. For those at the top, salaries soar into the $1.15 million-a-year bracket.

10 The creme de la creme of the fraternity are sponsored by companies such as Hang Ten and Quicksilver, whose somewhat radical motto reads: 'Admit nothing. Deny Everything. Make Counter Accusations'. The mantra chanted each night by any self-respecting surfer is: 'A winner will be the surfer who executes the most radical and powerful manoeuvres in the biggest waves with the greatest control and style for the longest distance'.

11 Many of the congregated were born too late to be considered genuine moon children of the '60s, but they did the best impressions they could. There were the odd few who could claim with validity that 'we were there', but in the '80s, talk about the 'Nam days' is doomed to sound like a long-playing record.

12 John Millard turned a sprightly-looking 50 recently. He saw more action than most during the war, commanding a 30-man ambush unit on the Cambodian border. It was the first time back for the emergency room physician, who acted as emergency medic for the competition.

13 'When I was first asked about coming here, I had to think for a moment,' he said. 'There were a couple of times I didn't think it was a good idea, but I had to get a feeling of what was going on here now. I was genuinely concerned about what happened to these people, some of whom hung on to the wheels of our planes as we pulled out'.

14 'It's so nice to have these people look you in the eye and smile, and to see the girls in bright pretty dresses instead of black uniforms. I'm surprised at the amount of freedom they have here under communism. I don't feel a heavy hand, unlike in Beijing'.

15 It wasn't a huge crowd of locals that turned out to watch the event; the largest gathering was on the first day when dancing dragons and firecrackers marked the opening. Those who did make the 15-kilometre journey from the town centre were as bemused as they were delighted by the antics of 32 foreigners jumping on and off pieces of plastic, and the accompanying culture they brought with them.

16 One little Vietnamese girl who tried particularly hard to enter into the spirit of things, spent the last day of the event saying: 'Totally awesome, man'.

Reproduced by permission of the South China Morning Post.
Appendix 2

A sample of interview extracts

Item 10: Why did the four Vietnamese take part in the Saigon Floating Hotel Surf Pro?

Answer: Because it is their beach/out of courtesy.

Subskill intended to be tested: making inferences

J1: E E
Students may not understand the whole part of the text, that is, 'Admittedly, the four Vietnamese taking part in the Saigon Floating Hotel Surf Pro, who were there courtesy of a wild card on the grounds that it was their beach, didn't pull off the manoeuvres as well as the 32 other professionals'. But they will still be able to answer the question, depending on what answer is expected from them.

J2: E E
The answer is quite straightforward 'except that our students may have lexical and language problems'.

J3: E E
'... they would have difficulty in finding out how these four Vietnamese are chosen but they should get the phrase 'it was their beach'.'

J4: E E
The answers are there. The phrase 'on the grounds' is used as a connective.

J5: E E
Students can guess the answer. Although some words like 'courtesy of a wild card' are difficult, students can still get the answer.

J6: E E
Students can copy directly from the passage.

J7: E E
If students understand the vocabulary, they will be able to get the answer. Our students could copy the relevant part from the text.

J8: Me Me
The words 'wild card' are used as a metaphor and are a key phrase to understanding the whole sentence.

J9: E E
The answer is 'it was their beach'. Again, the students could copy from the text.
Exploring the validity of ESL reading test items by means of judgemental procedures

Susanna Ho Pui-san

References


