An Analysis of a Japanese University Entrance Exam Reading Question:Iwate Prefectural University(1998)

Adrian Cohen

1. Theories regarding the testing of reading

The world of applied linguistics and the teaching of foreign languages has yet to produce one method of testing reading (or indeed any other linguistic skill) that can be said to be clearly superior to any other. Indeed, Alderson et al.(1995: 45) describe the prescription of specific test methods to the testing of specific skills as perhaps the "Holy Grail of language testing". It is important to bear this in mind when developing a test of reading and to recognise the inherent limits within any form of test that may ultimately be devised. In developing a reading test based on a particular passage, for example, we should note that the main thrust of Jafarpur's (1987) paper is that the use of such reading passages and questions in traditional reading tests is open to strong criticism. While Jafarpur may himself not have found the perfect answer to the question of how we should test reading in the short-context technique (see implied criticisms in Owen 1997), he does summarise at least seven objections to passage-based tests and backs them up with references to empirical research. Thus we should perhaps preface our justification of a particular passage and questions with a justification for performing this type of test in the first place.

In this context, it is interesting that Hughes (1989: 118-9) ignores any discussion of the advisability of basing reading tests on written passages and instead assumes that this will be the case, and offers advise on the types of texts that are appropriate and their selection. To an extent, this question is simply representative of the fundamental limitations of all linguistic enquiry, namely that we can only observe linguistic capabilities both indirectly and incompletely.

Bachman (1990: 10) identifies the worry as to "whether we can adequately reflect 'real-life language use' in language tests", but it does seem probable that, at the most basic level, we will be somehow testing students' reading ability by requiring them to read. Indeed, Jafarpur's primary objections are concerned with the length and authenticity of reading passages, not with the issue of presenting passages in the first place. The more common debate within the literature is in fact not concerned with the type of text to be presented, but with the type of questions that are set about the passage. This is a question that Jafarpur ignores completely and although he compares the short-context method with the cloze test, he enters into no formal discussion of the merits of different types of questions, and adopts the multiple-choice format without comment. While we need to take note of the arguments about text length and authenticity, it is also essential to consider in some detail the arguments concerning the type of tasks that students will be asked to perform.

Nevertheless, there is little point here in rehearsing once again all the arguments for and against the different methods of testing reading. Such methods are very well documented and the arguments for and against multiple choice, short answer, cloze (and its variants e.g. the C-test), selected gap filling and information transfer tests can be found for example in Hughes (1989: 120-129), Weir (1990: 43-51) and Alderson et al. (1995: 44-56) among others. Suffice it to say that all of these established techniques, while having individual limitations, can be justified in certain situations to the extent that of these three writers, only Weir is prepared to make an overall recommendation (short answer together with selective deletion gap filling). It is clear that the individual testing situation and its implications on validity together with the individual preferences of test writers will often as not be the deciding factor. This is not to condone the random selection of testing methods, but to accept that justifiable arguments can be made for any of the major methods in many testing situations.

2. The testing context and question choice

It is necessary, therefore, to consider the specific situation for which we are writing a test and for the purposes of this paper I have chosen for analysis the English entrance exam to my own school: Iwate Prefectural University. The test in question was the second of two opportunities for entrance and consisted of three reading comprehension/grammar questions the first of which was developed by myself. The other two questions (which will not be analysed here) consisted of passages followed by true/false questions, gap filling and translation from both Japanese to English and English to Japanese. The test was taken by 769 students applying for approximately 300 places in the Faculties of Nursing, Policy Studies and Social Welfare. Students were given one hour to answer all three questions. The question under analysis is presented in Appendix 1.

Development of this question was constrained by a number of factors that illustrate the practical problems of test development. Firstly, the use of multiple choice questions was forbidden by the school president who insisted that in all subjects students be made to produce, rather than passively choose responses. Secondly, the level of vocabulary and grammar was limited to the Ministry of Education guidelines related to the high school English syllabus. This includes lists of suitable vocabulary and grammatical forms for the level. Issues of content validity gave strong pressure for the format of a simplified reading passage based on an original text. Such passages represent the bulk of all approved high school texts and since Japanese high school students can be presumed to have had no other systematic exposure to English texts, there would be little justification in introducing authentic texts such as the newspaper advertisements used by Jafarpur. Indeed, the latter justifies his own short-context texts by the need to include "general topics of a nontechnical nature and with concepts familiar to, or required from, college students."

As a text, I chose to adapt an account of an experiment on chimpanzees described

by W. Köhler in his book *The Mentality of Apes* (1957). While the book itself was considered to be both too technical and difficult for the students, an adaptation was thought to be of general interest and was carefully screened for language of an inappropriate level. The test method was limited by the fact that initial estimates of candidate numbers suggested anything up to 10,000 candidates could be competing for the 300 places over two exams, although in fact the total figure was closer to 4,000, the bulk of whom took the first exam which is not discussed in this paper. Given the fact that the marking was to be conducted by the six English-teaching members of staff with support from six teachers of other foreign languages, it was considered inadvisable to produce extensive short-answer questions. A cloze test (and its siblings) was rejected because of the mechanical nature of the deletion process which would make some items irrecoverable by students (Alderson et al. 1995: 55) while an information transfer was considered to be unfamiliar and therefore inappropriate for Japanese students. This leaves selective deletion gap filling which suffers from the criticisms that it tests largely sentence bound reading skills (Weir: 48) and that it is very difficult to ensure that there is only one answer for each gap (Alderson et al.: 54). An attempt was made to reduce the effects of the first criticism by placing the gaps in a summary of the original text (as discussed in Hughes: 122-4) so that hopefully some general comprehension skills would be required, as well as an understanding of sentence level grammar and content. Nevertheless, an examination of the question shows that perhaps all items except for item 1 could be solved without reference to the original text, at least by a native speaker. This is partly the result of an attempt to address the second criticism so that the potential for multiple answers would be reduced. A choice of more content words would perhaps have amounted to a greater test of comprehension, but would have led to a far greater variation in possible answers. It should be pointed out that the original test specifications (which were unfortunately never formally written out) were not that the test should be of purely reading skills, and that a knowledge of sentence structure and lexical patterning were considered equally important as subjects for testing. The

number of possible answers was further reduced by supplying the first letter of the answers for items 2, 4, 5, 8 and 10.

3. Practical issues and problems

The nature of the Japanese university entrance exam system means that there are no opportunities for the pre-testing of materials. Tests are considered to be top secret documents, and if seen by anyone outside of the university staff before administration of the test, must be discarded. This creates enormous limitations on the reliability of tests and is a source of constant complaint by testing experts within the country (as noted for example by Hubbel et al. 1997). Thus the suitability of items can only be judged after the test has been administered, by an analysis of the candidates' responses. Such analysis is rarely in fact conducted and it is extremely rare for an item to be discarded at the marking stage unless a glaring mistake (for example in the wording of a question) is discovered. For the purposes of this paper I undertook an analysis of 132 student responses to the question.

Marking of the question under consideration was conducted by the twelve foreign language staff members working together in one room over a period of two days. An initial answer list was prepared but the marking orientation given to the supporting staff made it clear that alternative answers would inevitably appear and that all such variations were to be reported to the drafting committee of four. The committee would make a decision regarding the acceptability of each variation and the number of marks it was to be allocated, and this information would in turn be added to each marker's list. Three marks were awarded for a correct response while two marks were given for a minor spelling error and one mark for a grammatical error such as writing 'built' instead of 'building' for item one. The flexible nature of the marking scheme was considered to be the fairest way of ensuring that all candidates were given appropriate scores and the fact that the marking scheme was unified in real time helped reduce the possibility of

confusion. All marking was checked by a second marker and the calculation of points was conducted by a third and fourth marker.

4. Analysis of the question

An analysis of students' responses was attempted to assess both the suitability of the items in terms of level and the types of mistakes that they tended to cause. An analysis was undertaken which tried to record whether errors were of a grammatical, lexical, or spelling nature or if the item had been left blank. Obviously however, it is impossible to be categorical as to the source of candidate errors. Apart from the impossibility of knowing precisely what each candidate was thinking at the time of response, clearly there is a great deal of overlap between categories. Thus, for item 1, where the correct response was 'building' an answer of 'build' or 'built' (both of which were very common) would appear to be purely grammatical. However, a response of 'pick' could either be a pure guess (a word plucked from the text) or a genuine lexical and grammatical error (in other words the candidate has failed to grasp the meaning of the passage and also fails to notice the underlying grammatical construction of the sentence). As far as possible candidates were given the benefit of the doubt and thus 'pick' would be marked as both a grammatical and lexical error. Other problems of analysis are also apparent. For example, item 6 requires knowledge of the phrasal verb 'point out' and thus would appear to be a question focusing on lexical knowledge, yet the vast majority of incorrect responses involved grammatical words such as 'of' or 'that' which would seem to suggest an incomplete understanding of the way such grammatical words are used. Such responses were marked as grammatical errors but it is clearly possible to argue that they should be marked as lexical. Similarly, item 4 ('later') produced many responses of 'logic' which is grammatically inconsistent with the sentence structure, yet it seems likely that the response was a result of the word 'solution' appearing in the sentence and that combined with the prompt of the letter 'l' a lexical connection has been made between the words 'solution' and 'logic'. Here the lack of understanding of

meaning was considered stronger than the lack of understanding of grammar and the response was recorded as a lexical error. Given the problematic nature of such analysis, the value of such an exercise may well be called into doubt. However, given the inability to pre-test items on genuine populations of students, an attempt to gain some understanding of the nature of mistakes made by candidates can hopefully help the production of more appropriate tests in the future. Further analysis would allow a more comprehensive scheme to be devised so that real lessons can be gained, both about Japanese high school students' English and about the nature of the test itself.

5. Discussion

The first, and most obvious statistic that can be gleaned from such an analysis is the facility value for each item. However, given the fact that scores of two or one point were given for spelling and grammatical errors respectively, a simple record of the percentage of candidates supplying the expected response does not give us a full picture of results. Nevertheless, it does give us a crude view of the suitability of each item for the level of candidates and so such figures are presented in table 1 for a total of 132 candidates analysed.

Item No.	1	2	3	4	5	6	7	8	9	10
Expected Response	building	other	first	later	clear	out	of	solving	to	hung
F.V.(% of Expected responses)	20	94	51	11	3	18	55	43	98	26

Even given the limited nature of this analysis, it is clear that items 5 and perhaps 11 are far too difficult for this level of student and that items 2 and 9 are too easy. Perhaps we can congratulate Japanese high school teachers on their ability to teach the 'to infinitive'! A study of the data on grammatical and spelling errors

helps to illuminate some of the other items, but it is clear that students have more chance of picking up marks for what we could traditionally call 'open slots' such as items 1, 8 and 10 than they do for slots that are either lexically or grammatically more 'closed' such as items 3, 4 and 6. Although time has not yet allowed a full analysis of all the items in this way, we can report that 67% of students made a grammatical error in item 1, meaning that a total of 87% of candidates received at least one mark for this question. Similarly, 46% of students made grammatical errors in item 8 which allowed them to receive one point so that 89% of students received at least one point for this item. In contrast, there were no half measures for items 3, 4 and 6 meaning that students who failed to get the expected response or legitimate alternatives (such as 'once' for item 3) received no marks at all for these items. A most remarkable statistic revealed by this rudimentary classification concerns the number of spelling errors. Out of a total of 1320 possible responses there were only eleven spelling errors. This is particularly remarkable when it is considered that a student who misspelled 'built' for example, would be recorded as having made both a grammatical and a spelling error. We can perhaps attribute this lack of spelling errors partly to the fact that the items require only (relatively simple) single word answers, and that we would expect to see many more spelling errors in longer responses, but the lack of spelling mistakes certainly surprised this writer.

6. Conclusion

Unfortunately, further lessons to be drawn from this particular question will have to await further enquiry. However, we can now return to a more general consideration of the question as a whole. We can note that the rubric was clear in that it was understood by all students (no student left all responses blank or entered a response other than a single word), but this is hardly surprising given the fact that the rubric was in Japanese. We can also note that despite the practical limitations of the test environment there are elements of the question that could be greatly improved. Perhaps the most obvious of these is the need to create items

that rely more heavily on the reading of the passage as a whole. The fact that so many of the items could be answered without even looking at the reading passage suggests that the question is severely flawed and that as a reading test the main passage is almost superfluous. (Anecdotally however, it is interesting to note that during the test a quite large proportion of students were seen to be methodically translating the entire passage into Japanese before attempting to enter responses. This is clearly a problem of exam technique but also perhaps is a reflection on the type of task that students regularly encounter in high school). In addition, the analysis of individual items shows the problems inherent in any test that is not rigorously pre-tested and although there seems little chance that this is likely to change in the near future, the publication of such findings can perhaps gradually create an understanding of this fact. Clearly, some of these items were more suitable than others and an opportunity to pre-test the items would have led to a number of them being withdrawn and others modified. The classification of errors by their grammatical and lexical nature led to the realisation that deletions that are very limited in nature, either because of grammatical structure, or because of the strong concordances that are to be found between many lexical items, give students very little leeway and tend to produce responses that are either clearly correct or clearly wrong. While this may be desirable from the point of view of reliability, it is not clear that this gives us as accurate a picture of candidates' language ability as a whole, compared to questions that allow a more varied response. Clearly, this is leading towards an argument for short-response questions over the gap filling question used here, however there may be an argument that single deletions which allow a varied response offer a compromise between the difficulty of marking that afflicts short responses and the lack of linguistic information that can mar narrowly confined gap deletions.

Appendix 1

次のエッセイを読み、その内容と一致するように、エッセイの要約の空欄に適切な 1 語を入れなさい。ただし、空欄 2, 4, 5, 8, 10 には、指定された文字で始まる 1 語を入れること。

(Translation: Read the following essay and the summary that follows it. Fill in the blanks in the summary as appropriate to the essay. In numbers 2,4,5, 8 and 10 the answer should begin with the letter given.)

エッセイ (Essay)

Can chimpanzees build with boxes to help them reach a banana? A professor hung a banana from the ceiling. In the same room were placed two wooden boxes about four meters away from the banana but quite close to each other. When Sultan entered the room, he immediately picked up one box and placed it below the banana. He then stood on the box and seeing that he was not high enough, got down and picked up the other box. Although he waved the second box over the first box he did not place it on top. Instead, he became very angry and started running around the room. Then he suddenly stopped, as if the solution had just hit him. He ran up to the first box and placed the second box on top of it before climbing up and taking the banana. After this, Sultan always easily solved this problem, even building a tower of three boxes in order to get a banana, yet it is interesting that at first he could pick up the second box and fail to see the solution, even when holding the second box over the first. It seems that while chimpanzees can come to the same solution as humans, their minds do not see the problem in the same way, so that they cannot immediately see the logic of building.

(Adapted from W. Köhler: The Mentality of Apes, Pelican, 1957).

エッセイの要約 (Summary)

This essay is about chimpanzees [1](b) with boxes. It shows that chimpanzees can build to get a banana, but that they do not [2](s) the

problem the same as humans. Whereas humans can immediately see that the two boxes need to be placed on top of each [3](o), Sultan cannot at [4] (f) see this and only [5](l) does the solution become [6](c)to him. However, the essay points [7](o) that chimpanzees are capable of learning. Thus, in later experiments, Sultan had no problems in [8](s) this problem and later he was even able [9](t) build with three boxes when the banana was [10](h) higher.

References

Alderson, J. C., Clapham C. and Dianne Wall. (1995). *Language Test Construction and Evaluation*. Cambridge: C.U.P.

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing*. Oxford: O.U.P.

Hubbel, J., Pendergast, T., Thrasher, R., Yoshida, K. and K. Ohtomo. (1997) "Testing Colloquium". JALT International Conference 1997.

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: C.U.P.

Jafarpur, A. (1987). "The short-context technique: an alternative for testing reading comprehension." *Language Testing* 4,2.

Köhler, W. (1957). *The Mentality of Apes*. London: Pelican.

Owen, C. (1997) *Testing*. Birmingham University MA TEFL/TESL Course.

Weir, C. J. (1990). *Communicative Language Testing*. London: Prentice Hall.