Recommendations for the Hannan University English Entrance Exam

――阪南大学英語入学試験改善への提案――

Gordon G. Wilson
Yoko Wilson

Abstract

この論文は阪南大学の英語の入学試験に関しての批評と将来の試験改善についての実践的な提案からなる。現在，短所として次の点が挙げられる：ひとつの集合に属する学生の選抜に複数の試験が使用されている。試験的運用がない。試験項目が少ない。試験難易度がカット・ポイントよりも高い。空欄充当問題が問題文の早い部分に設置されている。空欄充当問題のまえに内容理解に関する問題が存在する。ひとつの文にふたつ設問がある。言語モードがテスト・モードと合致しない。

次の点が提案されている。

- 可能な場合，試験の数を減らす。
- テストを試験運用する。
- 答えの選択肢を減らし，試験項目を増やす。
- 読解文は長い文をひとつでなく，短めな文を複数使用する。
- 試験項目バンクを創り，良い項目は再利用する。
- 試験難易度をカット・ポイントに近づける。
- カット・ポイントに近い難易度の問題を増やす。
- 空欄充当問題を問題文の後の方にあてる。
- ひとつの文にふたつ以上の設問を作らない。
- 空欄充当問題を内容理解に関する問題の後ろに置かない。
- リーディングの問題にはテキスト・ベースの設問を，リスニングとスピーキングにはオーディオを使用した設問を使用する。またはオーディオ機材を使用しない際にはリスニングとスピーキング・モードの言語は避ける。

付け加えて，試験フォーマットの説明用「スタイル・シート」を製作し，試験製作者に配布することが提案されている。

This paper critiques the English element of a Hannan University entrance exam and offers practical suggestions for future tests to increase the quality. Weaknesses include: multiple tests are used to select from one pool of candidates, the lack of piloting the test, the small number of testing items, the test difficulty lies above the cut-point, item

completion questions early in the reading passages, item comprehension questions preceding completion questions, dual testing on individual sentences, and language and test mode inconsistencies. The following is recommended:

- If possible, reduce the number of tests.
- Pilot the tests.
- Reduce question options and increase the number of questions.
- Use few shorter passages rather than long ones.
- Create an item bank and recycle proven items.
- Aim the test difficulty closer to the cut-point.
- Increase questions with difficulty near the cut-point.
- Avoid completion items early in the passage.
- Avoid loading individual sentences with two or more questions.
- Avoid completion item before content questions.
- Use text based questions for reading items, audio for listening and speaking items. Or avoid listening and speaking language, if audio is not used.

In addition to these recommendations, the creation of a testing 'style sheet' which describes the target test format and can be distributed to test creators is suggested.

Introduction

Entrance exams are high stake tests which can influence the candidates' future lives. For this reason, fairness, efficiency, reliability and validity are extremely important. This paper will critique the English element of a Hannan University entrance exam and to offer practical suggestions for future tests to increase the quality.

We will critique the English element of the entrance exams of Hannan University, carried out in 2001. All the statistics, cited in this paper are public and came from Hannan How To Book, published by Hannan University as a guidebook for potential candidates. Hannan University is a relatively small private school (enrollment: 6,000), and like other schools, has a declining number of applicants as the population of 18-year old declines. In addition to declining enrollment, the following issues directly related to testing itself must also be taken into consideration. First, entrance examination fees are an important source of income for private schools and maintaining a base of applicants is a priority. Second, the difficulty of the test contributes to the public image of the school: a school with difficult entrance examinations tends to be considered as more prestigious. Other

functions of the exam include selection of candidates so as to include the highest possible level of students, and influencing the wash-back effect on the high school curriculum.

Before moving into the question-by-question analysis of the exam, we will give an overall critique of the entrance exam system of the school.

There are other criteria used to select candidates: Candidates' high school reports and Center Test scores. However, we will not cover these in this paper.

Hannan University has major three different kinds of entrance testing categories which function to increase the enrollment and applicants. (Though there are other entrance testing categories for foreign and adult students, we will not discuss them in this paper.) One of the three mainstream of testing categories is the *Suisen* tests in which the candidates are selected based on their high school report and/or the result of the entrance examination. The other two are called *Ippan A* and *Ippan B*, testing systems in which the candidates take the examinations and no other data is use in selection. We will focus on *Ippan A* exam because it is the testing category used by the largest number of the candidates. *Ippan A* consists of three tests, implemented on three sequential days (in 2001: February 6th, 7th and 8th.), resulting in problems with fairness.

Critique of format

The major weakness with *Ippan A* is that three different sets of tests are used to select from one pool of candidates. That is, it is almost impossible to fairly and accurately compare the ability of the students who take different tests. Students of one group, who have equal or even higher ability than candidates of the other two groups, possibly fail to pass. In other words, the evaluation of the candidates can be affected by receiving an easier or more difficult test. To avoid this problem, Item Response Theory (IRT) could be applied to compare the results of different tests, by estimating the difficulty of the test items. (McNamara 1996) However, application of IRT is unrealistic in the current context which excludes the possibility of pilot testing. In light of this, we will discuss other ways to improve the fairness of the tests later in this paper.

An *Ippan A* exam typically consists of two sections. One section is a reading test with one relatively long passage (a little less than one page) and questions about it. The other consists of various shorter question formats which are usually a combination of dialogue reading and grammar items. They are all multiple choice with five options. The

format of the questions is quite inconsistent among the three tests, aside from all being multiple-choice. For example, the very first parts of all the tests are questions based on dialogues between two people. However, the items from the exam on February 6th were completion questions by choosing one correct answer from the 5 options, while those on the Feb 7th and Feb 8th were to choose one correct answer which describes the place where the conversation took place or the relationship of the interlocutors.

This inconsistency can change the difficulty of each test and make the results less similar. Therefore, unless the difficulty of the items are adequately estimated using IRT, it is impossible to make the 3 tests equally difficult.

Critique of Individual Questions

Now we will take a closer at one of the tests. We will critique the exam, *Ippan A* given on the 2nd day.

The first part of the Feb 7th exam, the 7 items of Section A, (ア～キ) are tasks of reading short dialogues between 2 people and answering questions. (The original examination questions are in the appendix.)

First, giving conversations as reading passages should be avoided, because conversations are normally carried out as oral communication and not in written form. The language which the students are tested on, should be offered as authentic as possible and this should include consistency in the mode of communication. Therefore, if these conversations and the following questions had been given as listening comprehension, it would have created no problem with authenticity. It is possible, however, that the test creators' intention was to examine the candidates' reading comprehension of casual English. In this case, they should have used casual text-based language which is more authentic as reading, such as memos, Email messages or text-chat on the computer. However, the conversations themselves are very natural and the general focus on comprehension may result in a positive wash-back effect on high schools' curriculum to motivate the study of more communicative elements as opposed to discrete grammar elements which have been the main target of entrance exams in Japan.

Another weakness of this test is that some of the question instructions given in Japanese are ambiguously written. One problem may be that there is no clear instruction to select only one answer. For example, in question three: "When did the conversation take place?" some test takers may choose one out of the first two options, 'morning'

or 'evening' and another answer out of the rest, a season, intending to formulate such an answer 'spring, evening'. Though this is perhaps not very likely to happen, this point should have been stated explicitly. Another example of ambiguous instructions is the 2nd item, in which the test takers are asked the price of the movie tickets. The wording of the question in Japanese "How much does the movie fare come to?" is giving out a hint that the students are supposed to do addition, which rules out option four and five. Another weakness shown in this item is that the options are not presented in a numerical order. They should be presented in either descending or ascending numerical order, as Haladyna (1994) suggests, in order to reduce the applicants stress level and time. This point may seem rather trivial, but collectively it can result in an overall increase in the test time spent on the target. This can lead to possibility of being able to offer more questions in the equal amount of time, which results in higher reliability.

Another weakness is the low number of test items. This can result in the reliability of the exam being relatively low. One way of increasing the total number of questions is to cut the number of options from five to four. Because four options are more than enough, as creating four options often results in one of them being implausible (Brown 1996), cutting the number of the options should not hurt reliability. As an example, the sixth question would not have to have the second option 'These people went on a date.' which sticks out from the rest of options which all describe rather unpleasant situations.

The B section of the test consists of completion questions on conjunctions and prepositions. The total number of the questions being five this section is too small. To increase the validity, there should be 20 or more. Another weakness is that these questions seem to be grammatically too complex for the majority of the candidates. That is, assuming that an acquisition order actually exists (Ellis 1994), subjunctive past will be learned later on a very late stage of the order. Therefore this section may help distinguish higher-level candidates, but it may not be contributing to discriminate lower-level students where the cut-point may be most likely to fall. Especially with such a low number of questions, a difficulty level nearer the cut-point should take priority. However, if the goal of having such difficult questions on the test is to keep the public image of the school high by increasing the overall difficulty of the test, this question may be serving a function. It could be said that a test efficiently tuned to the appropriate level and cut-point of the students, and therefore improving the quality of the test and successful candidates, would contribute to the overall public image of the school.

In terms of format, the blank right at the beginning of a sentence, as in the third

question, should be avoided. (Haladyna 1994) This way, it is possible for the applicants to get the necessary information to fill in the blank and they will not need to read the sentence twice. This point also contributes to cutting down on the time spent on each test item.

The second part of the test consists of a reading comprehension section. A positive element of the passage is that it is somewhat entertaining and not too difficult. Therefore candidates may maintain interest in the passage itself. One overall weakness of this reading section may be that some questions require too much of inference. Particularly the fourth question is weak on this point: assessing what kind of person the judge was, although in the passage this remains ambiguous. This has more to do with individual interpretation of the passage and all the options seems to be more or less plausible. The fifth question also suffers from the same problem. Though the option three of question five does not seem to be as possible as the rest, the rest of the options all seem to be somewhat plausible. Although the candidates are asked to choose the best answer for this question, for there is no 'correct answer' questions of this type should be avoided (Brown 1994). Although these criticisms may seem minor to some, considering the number of questions, reliability of each item should be a priority.

The first few lines, or more preferably the first paragraph, of a test passage should not contain blanks, in order to allow the test takers to build a schema. Moreover, the third line contains two questions. This double testing on an individual sentence could result in one question providing a hint for the other question or missing one item could result in a lack of understanding, which causes the other item to be missed.

Also, it makes logical sense, to make the time students spent on this section more effective, the completion type questions (questions 8 through 12) should have been asked earlier in this section. This way test takers could have gone back to the completed passage when they solve the other questions.

Candidates with large amount of experience with American culture may miss the seventh question, which seems to be created as a vocabulary knowledge question. Some candidates, who have watched TV shows that take place in courtrooms, may have a hard time deciding between the first (a policeman), second (a guard) and fifth (an official) options, because they may assume that 'clerk' here may referrer to the people dressed like policemen or guards who serve as clerks in American courtrooms. Naturally, the closest word to 'a clerk' is 'an official', however those who actively used knowledge of

American courtrooms may have gotten this question wrong.

In addition, there should be more (shorter) passages rather than one long one to provide the test takers with 'fresh starts' which results in higher reliability.

For a thorough statistical analysis, more data is necessary. I analyze here statistical data available to the public. Assuming all the three sets of tests are of equal difficulty which is unlikely, but necessary for comparison, the average score of candidates for each faculty cluster around 50% (45.9-56.5%) shows that the difficulty was reasonable. However, the lowest score accepted for one of the departments being 13.6% is rather unreasonable, lower than a score generated randomly. However, since the candidates are required to take the test for another subject and the total score is used in selection, so this could be due to this student scoring high in the other subject. Even though the student who scored 13.6% may have shown very high ability in another subject which substitutes for his low performance in English, being able to get into the school with this score, which is less than pure chance (20%), renders the test meaningless. In this case, requiring candidates to take the test in one subject in which the student performs the best would be sufficient. This could be remedied by decreasing the overall difficulty of the test items, or increasing the items around the cut-point level. At any rate, this is a decision concerning the conflicting functions of increasing successful candidates and maintaining valid selection of candidates.

Recommendations

Based on the analysis above, we recommend the following.

First, one exam should be administrated on one day, instead of the current three different exams on three different days. In the case that this cannot be done, taking into account the test function of increasing the number of test takers, the following recommendation take on greater importance. Because there is no way to make all the three tests perfectly at the same difficulty level without using IRT, they consciously need to make the tests as similar as possible. Similar test formats and pilot testing the exam drafts as many people as possible, will increase validity, reliability and fairness.

Second, decreasing to four options from the current five option system would result in less time necessary to create the exams and enable more questions to be provided in the same amount of time. This would result in increasing the reliability of the test.

Adding more questions is recommended, because this exam has rather a small number of items. Particularly, the grammar section and the reading section should be expanded. There should be a chance for the candidates to make a fresh start, to eliminate the possibility that one weak topic reduces a students score.

To avoid invalid items, pilot testing is crucial. The best way is testing the items on a group of people who have close English ability to the candidate group. However, this would be almost impossible in Japan. Therefore, we recommend to pilot test items on Hannan university instructors or staff who have similar English ability to the candidates and have not seen the items before. Having the other instructors who did not create the items is naturally a plausible way to improve the items, test makers have English ability far too high to judge validity and are too close to the test. Also, asking pilot testers to 'think aloud' into a recording device as they solve the questions would provide additional insight into potential problems, which test makers cannot predict. This way, it will be possible to cut items with ambiguous instructions, etc. One down side of this method is that going over the recordings will take quite a lot of time.

We also recommend creating an item bank, a collection of test items that have been proven to be valid and functional. (Henning 1987) The test makers can choose items given in the past and include them in subsequent tests. Also, it is possible that they collect some functional items from end-semester tests. Though they are Criterion Referenced Tests while entrance exams are Norm Referenced Tests, some items must be applicable. One point to keep in mind when using items from the bank is that they alter small details, such as names, times. It should also be noted that recycled items should be introduced independent of the items with which they previously occurred and the order of distracters should be changed. Otherwise, some candidates may memorize the answers of the test items given in the past and the test result may simply turn out as reflection of their rote memory.

Another point that may require an alteration is the difficulty. It is crucial to carry out a statistical analysis to assess distribution. There is a typical pattern of candidates' behavior: they choose to go to the school highest in prestige among those whose tests they passed. Therefore, though the average scores are around 50, it is possible candidates who scored higher on these exams chose to go another school. Thus, the level of the candidates who actually get into Hannan can be lower than the average of all candidates, creating a false cut-point. Also, it must be noted that a major function of the entrance exam is discriminating between the candidates on the low end of the accepted

group and not stronger candidates.

Supporting data for decreasing the difficulty level is that the lowest score accepted for one of the departments being 13.6%, while maintaining a wide range of difficulty in the questions to serve the prestige function of the test, we recommend increasing the number of questions at the lower end. Increased validity at the range of difficulty nearer the cut-point will result in a smaller number of substandard students slipping through and also decrease the number of student with an adequate level who do not pass. This will result in a small but important increase in the overall ability level of students who enter Hannan.

Conclusion

This paper covered only a part of selection process at Hannan University. The English tests that I didn't cover, *Ippan B*- and *Suisen* tests appear to have similar problems to *Ippan A*. The recommendations in this paper are likely applicable to the other two tests. These recommendations are summarized as follows:

- If possible, reduce the number of tests.
- Pilot the tests.
- Reduce question options and increase the number of questions.
- Use few shorter passages rather than long ones.
- Create an item bank and recycle proven items.
- Aim the test difficulty closer to the cut-point.
- Increase questions with difficulty near the cut-point.
- Avoid completion items early in the passage.
- Avoid loading individual sentences with two or more questions.
- Avoid completion item before content questions.
- Use text based questions for reading items, audio for listening and speaking items. Or avoid listening and speaking language, if audio is not used.

The creation of a testing 'style sheet' which describes the target test format and can be distributed to test creators would aid in the consistent and successful implementation of the principles which are adopted by the test creation group.

References

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.
Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford, UK: Oxford University Press.
Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum.
Hannan University, *Hannan How To Book 2002年阪南大学入試ガイド*
Henning, G. (1987). *A Guide to Language Testing*. New York: Newbury/Harper.
McNamara, T. (1996). *Measuring Second Language Performance*. Essex, UK: Addison Welsley Longman.

Appendix (*Ippan A*: February 7th 2001)

- The format may be slightly different from the original, but the content is the same.

I 問いA, Bに答えよ。

A ア～キの設問に答えよ。

ア 次の会話が行われた場所を選びなさい。

A: Good afternoon, sir. How may I help you?
B: I would like to book a flight to Canada.
A: Will that be one-way or round trip?
B: Round trip, please.

1 A travel agency.
2 A library.
3 A police department.
4 A restaurant.
5 A car rental shop.

イ 映画代はいくらになりますか？

A: Two adults and one student ticket, please.
B: That is $7.50 each for the adults and students are $5.00.
A: Thank you. Here you go.

1 $12.50.
2 $20.00.
3 $17.50.
4 $7.50.
5 $5.00.

ウ この会話が行われたのはいつですか。

A: I am sure glad it's finally warming up.
B: Me, too. I forgot how nice it is to open the windows.
A: But soon it will be hot to keep the windows open.
B: Yes, but I can enjoy it for now.

1 Morning.
2 Evening.

3 Winter.
4 Fall.
5 Spring.

エ 二人は何をしていますか？

A: Hello.
B: Hello. This is Karen.
A: Oh, Karen. How have you been? I haven't heard from you in some time.
B: Yes, I have been rather busy lately.

1 Talking at a party.
2 Meeting on the street.
3 Watching television.
4 Talking on the telephone.
5 Eating lunch together.

オ 二人は何について話していますか？

A: I really hope I can get into Dartmouth.
B: I hear that it is rather expensive.
A: I can get a student loan if my grade stays high.
B: Good luck.

1 Getting into college.
2 A severe injury.
3 Travel.
4 Restaurant reservation.
5 Graduation.

カ 次の会話の前に起こったことを選びなさい。

A: I only turned my back for a moment.
B: Did you see anyone near your bag before it disappeared?
A: Yes, now that you mentioned it. There was a man there.
B: What did he look like?

1 There was an argument.
2 These people went on a date.
3 Something was stolen.
4 There was a flight.
5 Someone went missing.

キ 二人は何をしているところですか？

A: How hungry are you? Would you like to see the movie first?
B: That would be fine. I had a snack about an hour ago.
A: Great. We can decide where to eat after show.

1 Cooking dinner.
2 Watching television.
3 Doing homework.
4 Going on a date.
5 Waiting for their friend.

B 次の(1)～(5)の文が同じ内容になるように，(ク)～(シ)に入る最適な語を1～8から選び，その番号を. ～. にマークせよ。ただし，文頭に来る語も小文字である。

(1) I could not go out because (ク) the heavy rain.
(2) It was raining (ケ) heavily that I could not go out.
(3) (コ) it had rained heavily, I could not go out.
(4) The heavy rain prevented me (サ) going out.
(5) (シ) it had not rained so heavily, I could have gone out.

1 as 2 at 3 if 4 of 5 from 6 on 7 and 8 so

II 次の文を読み，設問ア～ソの最適な答えを，それぞれ①～⑤から選びその番号を . ～. にマークせよ。

There once was a nice old lady called Mabel Fenster. She was bored with her life and she decided that she wanted to do something exciting. She (a) to go to the supermarket and (1)steal some things.

She went to the (b) 24 hour supermarket and picked up a basket. She walked slowly around looking at the fruit and vegetables, frozen and canned food. When she thought no-one could see her, she quickly put a can of whole tomatoes into her coat pocket. A security guard who was watching the camera monitors saw her action and she was (c).

A few days later she had to go in front of a judge. He looked down at the little old woman.

'Miss Fenster, is this the first time you have done this?' he said.

Miss Fenster nodded.

'Well,' said the judge. 'That's too bad, but you are (d) enough to know better. Bring me the can of tomatoes.'

The judge ordered his (2)clerk to open the can and to (e) the tomatoes. There were six.

'Now, listen.' said the judge to Miss Fenster. 'I am going to send you to prison for six weeks. One week for each tomato.'

Miss Fenster looked up at the judge and said cheerfully: 'Oh! I'm glad I put the can of peas back…'

ア For how long was Mabel Fenster sent to prison?
① 24 hours. ② 60 days. ③ a few days. ④ 6 weeks. ⑤ 1 month.

イ What did Miss Fenster steal?
① Fruit and vegetables. ② Canned tomatoes. ③ Frozen fish.
④ Some peas. ⑤ Nothing.

ウ Where did she put the stolen goods?
① In her pocket. ② In a basket. ③ In a can.
④ Under her coat. ⑤ In a shopping bag.

エ The judge was very…
①…friendly. ②…understanding. ③…helpful.
④…kind. ⑤…severe.

オ Why was Miss Fenster cheerful?
① She liked the judge. ② She thought the judge was fair.

③ She wanted to go to prison. ④ Her punishment could have been worse.
⑤ She enjoyed stealing from shops.

カ Underlined (1) means…
① look at. ② buy. ③ test. ④ borrow. ⑤ take.

キ Underlined (2) means…
① a policeman. ② a guard. ③ a detective.
④ a shop manager. ⑤ an official.

ク Which word best fills in the blank (a) ?
① promised. ② expected. ③ decided. ④ managed. ⑤ intended.

ケ Which word best fills in the blank (b) ?
① neighborhood ② official ③ closed-down
④ noisy ⑤ risky

コ Which word best fills in the blank (c) ?
① sent home ② surprised ③ caught
④ pleased ⑤ admired

サ Which word best fills in the blank (d) ?
① poor ② rich ③ bored ④ old ⑤ fast

シ Which word best fills in the blank (e) ?
① taste ② count ③ slice ④ juice ⑤ eat

ス Which of these sentences is *not* true?
① Miss Fenster almost took some peas from the supermarket.
② Miss Fenster wanted to do something exciting.
③ Miss Fenster was a very bad old woman.
④ Miss Fenster was seen taking some tomatoes.
⑤ Miss Fenster had never taken anything before.

セ Which of these sentences is *not* true?
① Miss Fenster wasn't worried about going to prison.
② Miss Fenster put the peas back.
③ Miss Fenster said she had done nothing.
④ The judge decided the punishment by the number of tomatoes.
⑤ The judge didn't care that Miss Fenster was an old lady.

ソ Which of these sentences is *true*?
① The security cameras were broken.
② There are six peas in a can.
③ Miss Fenster didn't have to go to prison.
④ Miss Fenster thought no-one was watching her.
⑤ The can of peas was brought to the judge.

(2001年12月13日受理)