言語センター広報 *Language Studies* 第12号(2004.1)小樽商科大学言語センター

Development and validation of a doctor-patient role play test for medical students in Japan

Ian Munby

[1. Introduction. History, purpose, content, and organization of the course]

In 1996, the Hokkaido University School of Medicine in Sapporo, Japan, requested the NES instructors of English in the neighboring Department of Language and Culture to develop a medical English course for its third and fourth year students in response to growing demand for English skills and knowledge of medical English in the international medical community (Holst and Evans 2000). The main goal of the course was to prepare students for careers in medical research wherein effective participation in international medical conferences was considered especially important. The resulting focus was therefore on English communication skills (listening, speaking and presentation skills) and the development of health and medical related vocabulary.

A secondary (minor) goal of the course was to prepare students for consultations with a small but steadily increasing number of English speaking patients in Japan, leading to the development of a doctor-patient role play test, the development and validation of which is the subject of this paper.

98 students were divided into four classes held concurrently by a team including the author and three other male NES teachers hereafter referred to as MH, RE, and BZ. None of the team has any specialized medical knowledge or training. Ten to thirteen 90 minute classes per semester are held for three semesters, two in third year and one in fourth year. The focus of the second part (semester) of the course is on illnesses and diseases.

In brief, students are tested in pairs at the end of the course. One student takes the role of patient with a role card detailing reason for visiting doctor, history, symptoms and lifestyle while the other takes the role of doctor. They perform a five-minute consultation and then switch roles and repeat with new role cards.

A total of seven diseases are studied in the course: arthritis, coronary artery disease (CAD), migraine, stroke, eating disorders (anorexia and bulimia), high blood pressure (HBP), and asthma, each representing one unit, or one ninety-minute session. Candidates must be prepared to perform a consultation with a patient suffering from any one of the above conditions without prior notice. Input of relevant data about each disease is provided by the Time/Life medical series video *At the time of diagnosis*. Copies of these videos are studied for homework by groups of four or five students who complete viewing tasks in their course books. In addition, students study a list of key words and phrases related to the disease and are tested on ten items at the beginning of each class. Furthermore, for each unit (disease) half the students in each class are required to write a doctor-patient dialogue while the other half write a patient role card (including notes about lifestyle and symptoms) of the type that will be used in the test. Class halves alternate each week. These written dialogues and role play cards are used in class for practice and collected at the end, the former for correcting and grading and the latter as a resource to assist in the design of role cards for the test.

[2. Test design]

What kind of test is it and what inferences can be drawn from the results?

This is a performance-based final achievement test where candidates are tested directly on their ability to produce spoken language to complete a communicative task in which they are required to demonstrate course content mastery. While it must be emphasized that this is primarily a speaking test in medical context, not a screening mechanism to determine whether or not test-takers may enter the workplace (McNamara 1996. 92), "indirect" predictive inferences can be drawn "about what candidates should be able to do in that real life context" (Weir 1993. 31). This reflects the dual purposes of the course which includes the need for students both to "manage the communicative and language demands ..of the criterion situation" (McNamara 2000. 49) and to develop control of medical English.

In order to construct such a test of spoken language, Weir (1993. 30) suggests first considering available theories on what is involved in the speaking skill and, second, defining the operations, performance conditions, and expected quality of output.

Operations

Bygate, cited in Weir (1993. 31), suggests that speakers draw on a repertoire of routines in order to communicate and a logical point of departure in test construction would therefore be the identification of the routines enacted to complete the type of task being tested. Weir (1993. 30) notes that operations involve both informational and interactional routines and improvisational skills. In this test, informational routines concern ways in which a doctor may present the symptoms, diagnosis and treatment of diseases to the patient and involve a considerable amount of factual information. Regarding interactional routines, asking questions relevant to diagnosis and responding to answers given are, for example, essential to medical consultations. In support of these informational and interactional routines, students are expected to perform evaluative routines as in the following examples taken from Weir (1993. 32).

Drawing of conclusions: diagnosing the patient's ailment
Explanations: explaining test results, the illness, and treatment to the patient.
Justification: explaining why a particular treatment option is advisable.

The term "consultation pattern" is used by the test developers to refer to the sum of all the identifiable operations involved in medical consultations. The following description of the six stages also appears in the course plan in the course book.

1. Greetings and introduction,
2. Patient presents complaint
3. Questioning about symptoms, history, and lifestyle
4. Physical examination, or making the diagnosis

5. Presenting the diagnosis (explaining the illness)
6. Explaining treatment and ending the consultation.

The consultation pattern also reflects the structure of the videos for each disease studied by the class and is reinforced through written dialogue training. The above stages are based on logic rather than extensive domain sampling of the kind used in the OET (Occupational English Test) as an analytical tool in test design. However, the same elements were observed in the OET (McNamara 1996. 104).

In the event of breakdown in the interaction, what Weir (1993. 30), drawing on Bygate (1987) describes as "improvisational skills" are enacted and involve "negotiation of meaning" and "management of interaction". As Weir (1993. 33) points out, the fact that this test consists mainly of interactions between two candidates, with the examiner playing only the role of nurse, it is not possible to "operationalize the testing of this ability".

In order to make the students aware of how their performance was going to be assessed it was decided to include a description of an excellent and a failing performance in the grading of the test in the course plan.

Performance conditions

According to Weir (1993. 39), conditions affecting performance on the test include processing under normal time constraints, degree of reciprocity/participation in developing interaction, purpose, nature of interlocutors, setting, role, topic, channel, and input dimensions.

Processing under normal time constraints Recorded samples of students performance in class, which were used for standardization, suggested that a time limit of five minutes per role play was appropriate.

Degree of reciprocity/participation in developing interaction. Although speaking rights are shared by both doctor and patient, the interaction is clearly doctor-driven and it is he or she who has the responsibility of "moving" the interaction through the key stages detailed earlier. However, it was thought that any imbalance in reciprocity conditions which may affect performance can be redressed through the exchanging of roles.

Purpose. The purposes of the test, discussed and agreed by members of the team, are to measure the students':

1. course content mastery, or medical knowledge, and related English terminology, of the illnesses studied on the course.
2. ability to interact effectively in English with a patient to diagnose a disease in a real-life-like manner using the consultation pattern.
3. speaking skills and level of communicative competence in English.

Since achieving realism is therefore a crucial issue in validating the construct of the test, changes were made to deal with authenticity-related problems which the team experienced in the preceding year's test (December 2000). Furthermore, it was decided to include the following details on the patient's role cards which had formerly been absent: name, sex, age, occupation, height and weight. The doctor was also supplied with the same information to increase authenticity and the patient's reason for visiting doctor was highlighted under a separate heading at the top of the patient role card.

A second, more serious problem experienced in the December 2000 test arose from the doctor's control of medical test results in diagnosis. Basically, the freedom of students in doctor role to produce their own test results allowed them an unwelcome freedom to diagnose whatever disease they wanted to diagnose. For example, a patient with appetite loss, supposed to be suffering from anorexia, could be diagnosed with ulcers (a disease removed from the syllabus this year) if the doctor preferred.

To counteract the problem, RE came up with a truly inspired idea. He suggested that the examiner took the role of nurse in the interaction, thus taking charge of producing results of physical examinations requested by the doctor. In this way, the examiner is able to direct, or redirect, the course of the consultation. It also satisfies a need felt by some for examiner involvement in the interaction. To meet the needs of the examiner in nurse role, potential results of medical tests are listed at the bottom of the examiner's copy of the doctor and patient role cards under the heading "nurse role card".

Interlocutors. Weir (1993. 39) suggests that the following conditions concerning the interlocutors may affect performance and, by extension, validity: the number of participants in the interaction, their status, familiarity, gender, and role. First, regarding number of participants, it was decided to adopt a format with two peers interacting in the roles of doctor and patient as opposed to one examiner as patient and a candidate as doctor.

There are clear advantages to the latter format wherein the examiner can test the doctor's improvisational skills by feigning misunderstanding, for example, and test medical knowledge and associated routines by asking for explanations. It would also reflect the original purpose of the course, to develop the requisite skills for conducting consultations with NES patients. Furthermore, a more level playing field could be established if the examiner's input and performance as patient remains more or less standard. In contrast, in the adopted two-peer system, a weaker candidate in patient role may affect the performance of a candidate with superior language skills in doctor role, or vice versa, a significant performance condition variable noted by Hughes (1989. 107) and by Weir (1993. 38) under the subheading of "input dimensions".

The following disadvantages in an interactional format where the examiner takes a patient's role should also be considered. First, the examiner's attentional capacity for effective rating would likely be constrained by having to concentrate on participation in the interaction. Second, as Weir (1993. 37) points out, the candidate may find it easier to speak to a peer, especially since it reflects the conditions in which the students were prepared for the test.

Regarding *status and familiarity*, the candidates share equal status as third year medical students and are known to each other. Careful scheduling ensured that the students were not tested by their own teachers to eliminate rater prejudice. Regarding gender, female students are in a clear minority but were not paired together to avoid "the gender effect" described by Weir (1993. 37). Finally with regard to *setting and channel*, students are left to imagine that the classroom or lecture hall represents the consulting room, and the channel is face to face, reflecting real-life conditions.

Quality of output

Quality of output concerns "the expected level of performance in terms of various relevant criteria". (Weir 1993. 30). The assessment criteria, or "band descriptors for grading", and the mark scheme, labeled "grading grid" were discussed by the team in two sessions prior to the testing day. In previous years, there had been no band descriptors as a guide to producing raw numerical scores, nor standardization of scoring, and students were rated purely on examiner intuition.

With reference to the purpose of the test, the criteria were divided into two categories, the test-taker's performance as a doctor and spoken language performance. The former was further subdivided into the following areas. The first of these is use of consultation pattern (CP) where the candidate is required to demonstrate control of the informational and interactional routines detailed earlier, as noted by Weir (1993. 41). The second is medical knowledge (Med) where the candidate is required to demonstrate that he has retained knowledge of the symptoms, tests for diagnosis, and treatment of a sample disease. The latter category, spoken language performance, was subdivided into two areas, fluency (F), relating to "smoothness of execution" (Weir 1993. 42) and grammatical accuracy (A) which concerns intelligibility and the degree to which the test-takers' control of the language system affect communication.

The third sub-category is manner (M), essentially a measure of pragmatic competence (Weir 1993. 42) or degree of doctor-like-ness.. The team members agreed that it was an important aspect affecting performance since a considerable number of seemingly inappropriate "speech acts" had emerged in the written dialogues. These included, for example, statements like "You will die, OK?" While McNamara (1996. 79) states that any decision to exclude socio-cultural competence from performance measurement would be "odd", given that Hymes' argument for its importance has been "universally accepted", the extent to which it can be reliably measured is questionable. According to results of my own research into NES teacher rating of appropriacy and inappropriacy in student output, judgements as to whether or not items were appropriate, and the precise level of pragmatic violation on a rating scale of 1-5, were markedly varied. In this sense, from a theoretical perspective, rating manner, or degree of doctor-likeness, is a dangerous exercise, especially without evidence, in the form of data from domain sampling, of what doctors actually say in similar consultation situations.

The rater's task was to assess the student's performance in each of the five analytical scales as excellent (8-10), moderate (5-7) or failing (2-4). Descriptions of what constituted an excellent performance and a failing one were included in the examiners band descriptors. Some of the language used to describe performance, for example "communicates readily" in the description of an excellent performance in fluency, were borrowed from the band descriptors of the IELTS speaking module. Others were agreed on in the course of team discussion, for example "runs risk of making patient feel ill at ease" to describe failing performance in the "manner" criterion. Additional uncertainty reigns here since it might also be argued that doctors would find it inappropriate to "put patients at ease" if their conditions were life-threatening and caused by unhealthy lifestyle.

Moderate performance constituted performances which were neither excellent nor failing. The mark scheme was designed to combine assessments of the students' performances on these five analytical scales into a raw numerical score adopting a system similar to the one described in the University of Reading Distance MA TEFL Student Handbook. For example, "moderate performance in most areas, perhaps failing in others" produces a score of 6 out of 10.

Test power

For course final assessment, students are given a grade with a weighting of 25% for each of the following components: vocabulary (average test score), written dialogues (average score), group presentations (about a disease of their choice not previously studied), and the doctor-patient role play test. Although it is not a career-determining "high stakes test" (McNamara 2000. 49), if combined scores from the above components fall below 50%, and the student fails in two other (non-English) medical courses, he or she must repeat the whole year of study. In this sense, test-designers have a moral responsibility to measure performance fairly.

[3. Empirical Validation]

What follows is a discussion of the validity, reliability and practicality of the test and examines evidence of how effectively it measures what it was designed to measure. Empirical evidence used includes transcripts of sample performances, data from standardization sessions, re-ratings of recorded performances, analysis of scores, or the "data matrix" (McNamara 2000. 56), recordings of discussions by the team and questionnaires completed by examiners and test-takers.

Validity

A good, or valid, language test could be found to have, according to Weir et al (1999. A9), the following four kinds of validity: construct validity, content validity, face and washback validity, and criterion related and predictive validity. The latter measure is unfortunately unavailable.

Adopting a view suggested by McNamara (1996. 16), one could identify two separate aspects of *construct validity*: *how* and *how well*. The issue of *how* involves an evaluation of the decision-making process in test design, or "a priori" validity, which has been discussed in the previous section. As for *how well*, we need first to examine samples of recordings and transcriptions of student output and score sheets to determine whether the operations the test is designed to measure are in evidence.

Analysis of the transcript of one performance, shows that the student follows the consultation pattern. For example, with reference to the routines detailed earlier, she draws conclusions in this way: "I think that you are suffering from migraine". She explains the illness; "*migraine is caused by .. um ..the blood vessels in your brain and get expand and it pushes the trigeminal nerve ..it's in your brain...". Furthermore, she justifies the recommended treatment option "*..and the beta-blockers is to prevent the pain. So you said that you lose part of your sight before you get headaches. It means that you're going to have headaches so if you lose part of your sight you have to take this beta-blockers to prevent the headache". However, it has to be recognized that detailed analysis of recordings of all 98 candidates would have to be made to justify any conclusions on construct validity in this regard.

Regarding *performance conditions*, evidence from recordings of 58 performances showed that the majority of candidates either finished or appeared to reach the final stages of the consultation and 49 candidates successfully diagnosed the patient's condition, with only 9 failing to do so.

Next, concerning *quality of output*, comes the question of whether or not the abilities the test was designed to measure are separately identifiable and measurable. In other words, the issue of whether or not there is statistical evidence of differing levels of performance, or competence, in each of the five analytical scales.

One candidate scores an exceptionally varied 10 for fluency, 8 for accuracy, 7 for use of consultation pattern, 6 for manner and 4 for medical knowledge, producing a final score of 7. However, only in a small minority of my own score sheets and those collected from MH and BZ, was "unitary performance" observed with equal scores on all scales, perhaps validating the distinction, or separateness, of the constructs. Unfortunately, RE did not check any boxes in the analytical scales on the score sheets dismissing them as an unhelpful distraction.

However, the validity of one analytical scale, namely manner, needs to be discussed again.. On the one hand it may be possible to identify pragmatic failure in some cases, taking an example from one performance where the doctor says "(your) blood pressure is terrible", perhaps qualifying as "running the risk of making patient feel ill at ease". On the other, regarding "excellent manner", statements like "*I'm very sorry to say this but I say you to stop this" spoken to a migraine sufferer who regularly partakes of two of its key triggers, red wine and cheese, appear to satisfy the criteria for high pragmatic competence at this level.

Some problems remain. First, while student performance never failed to provide something measurable for each of the other four analytical scales, manner was sometimes "missing" or at best conveyed through elements of performance which were hard to attribute to any specific speech act. In other words, some candidates sounded very doctor-like, but in what way it was impossible to tell. This could be viewed as a positive feature of performance measurement. However, one candidate was adjudged to have failed in all areas except manner, where she was rated as moderate because neither was there any sign of excellent or failing performance, nor did there appear to be anything upon which to judge her level of pragmatic competence at all. This problem of descriptors including "features that do not occur in actual performances" was noticed by Upshur and Turner (1995. 6).

Thirteen students were adjudged to have failed the test with a score of 4. According to evidence from examiner score sheets, the main cause of test failure was substandard fluency. However, since no candidate was awarded a score of less than 4, the validity of the range of available scores on the rating scales can be called into question. MH admitted that he didn't dare rate a candidate at less than 4, and I felt the same pressure. As to the possible causes of this phenomenon, which produces the effect of "cutting off the left edge of the bell curve", it appears that the examiners are unwilling to exercise the potential power of the test, by awarding 2 points for example, perhaps in recognition of the possible consequences for the student. This is possibly because it was considered that students entering the course with poor speaking skills, usually characterized by failing levels of fluency, were not to blame for their predicament, and that such a short course could not rectify the situation.

With regard to *content validity*, since this is an achievement test, the issue of whether or not the test effectively samples what has been taught is a key consideration. One complaint voiced in questionnaires completed by students is that some diseases, and therefore tasks, may have been more difficult than others. While RE admits this is probably true, MH writes; "The key point of the test was to see how well they explained things in clear English, so even if they made a mistake with the diagnosis, this would be largely irrelevant to our understanding of their English ability". This is mysterious since I thought we had agreed that the students were being rated not only for their English ability but also for their medical knowledge. Lack of relevant medical knowledge, though not relevant to their English ability, certainly does affect performance.

One candidate's performance and final score was, to my mind, affected by lack of medical knowledge of stroke. The average score on this disease was the lowest of the seven, with the highest score of 8 for this disease being awarded to a candidate who even failed to diagnose it, and was later unofficially re-rated at 7. Certainly the number of tests required to diagnose stroke are considerably more numerous than in any other disease. Strangely, one student complained that Anorexia was difficult because it had "very little information", a point on which I don't agree. As RE commented, degree of disease difficulty depends on the person and is a subjective matter.

Since this also a test of spoken English, it also needs "to be based on a theoretical model of whatever spoken interaction consists of" (Weir et al 1999 A11). As mentioned earlier, the pair format prevents the operationalization of improvisational skills, a major shortcoming since it is difficult to know under which analytical scale to give credit to students performing well in this area. For example, one candidate negotiates the meaning of her question when the patient misunderstands.

DOCTOR What kind of pain do you have in your head?
PATIENT Oh?

DOCTOR Is it both sides? Or .....
(PATIENT DOESN'T ANSWER)
DOCTOR Do you have the pain in both sides of the head or only one side of the head?
PATIENT Right side only.

In contrast, a student with poor improvisational skills may "escape unnoticed" in a role play where there was no communication breakdown to force the candidate to draw on them, while another with skills in this area may go unrewarded. This relates to another peculiar feature of this, the first test of speaking that I have ever seen that does not seek to measure pronunciation. As MH once said, we should not assess pronunciation if we don't teach it, and we don't. However, I noticed on many occasions that poor pronunciation prevented me from decoding what was said even though the students appeared to understand each other. For example, with one role play, I had to listen to the tape several times to decode the word "blurred" but there was no communication breakdown between patient and doctor. This is what happens in the monolingual class and in a test format where a native speaker is not one of the interlocutors.

With regard to *face validity* Weir (1999. A13) reminds us of two important points. First, face validity is not validity at all, and the same applies to wash-back validity. Second, test-takers may not feel the test is testing or measuring their abilities adequately and not bother to prepare for it, a serious concern in an achievement test which partly measures course content mastery. After all, why prepare for a test that appears to be nothing more than an elaborate lottery? Evidence from test-takers about their feelings towards the test is an important source of information here. However, it has to be admitted that only 35 completed questionnaires were received from 98 candidates, many of whom clearly had better things to do than fill them in, like study for a microbiology examination held later the same day, one of many complaints received.

Nevertheless, the overall impression was positive, and 83% of those who completed the questionnaire studied for the test, 62% believed the course prepared them well for it, 77% were satisfied with the time limit, and 68% felt the test was fair. In answer to question 6, "what do you think the test was really testing?" the comments elicited suggest that most candidates understood the aims of the examination, and 60% thought that the test was testing the required abilities well. In this sense, it could be concluded that in the opinion of the majority of a sample of test-takers, it was a valid test. However, three test-takers felt their performance was affected by the ability of their partners, raising questions once again about the pair format.

As for *washback validity*, which concerns the "extent to which the test has a beneficial or other effect on the teaching and learning which goes on in preparing for it" (Weir et al 1999. A13) assessment of the consequences of the test must also be considered. From a theoretical point of view, negative consequences are assumed to be suffered most severely by indirect tests where test formats are only indirectly related to target performance (McNamara 1996. 23), and since this is a direct test of a real-life skill which is practiced in class on a weekly basis, there wouldn't appear to be a problem here. Moreover, evidence from questionnaires completed by students on how they prepared for the test shows that most reviewed information on the diseases covered in the course, which hopefully had a beneficial effect on their learning.

Reliability

As Weir et al (1999. A5) point out: "if the numbers which summarize the test-takers performance are not consistent ... the test is fundamentally flawed". Reliable numerical evaluation, or scorer reliability, is especially important in this test for two reasons. First, spoken English is notoriously difficult to evaluate, even more so than written English. Second, since candidates are rated by different members of a team, standardization of scoring is crucial.

A thirty-minute standardization session was conducted on 30th. November, 2001. The four members of the team listened to recordings of two role plays recorded in class and rated the performance of the two doctors. The ratings balloted by each member indicate that the potential for wholly unreliable rating on test day the following week was huge. For example, in the second role play, BZ rated the candidate at 8 (excellent), while MH rated her as 4 (fail), prompting RE to remark: "I can't believe you failed her". Neither could I. I find it even harder to believe that MH considered the standardization session to be "OK" and that RE felt it was "good". If it was "OK" or "good" one wonders how much worse it would have to be to qualify as bad. This danger of scorer unreliability was also noted by two candidates in the questionnaires in answer to question 4 "Do you think the test was fair?" and may be related to the opinion of another that no student should be failed in the test.

On a brighter note, data from re-ratings of student performances yield more positive, but nonetheless unsatisfactory results. Regarding intra-rater reliability, re-ratings of my scoring of 32 candidates showed that while I assigned the same score to only 11 students, I was one point out with 16, two points out with 4, and three points adrift with one student. This represents a correlation of 0.76, which, while lower than the figure of 0.8 preferred by Upshur and Turner (1995. 9), is probably reasonable. As for inter-rater reliability, my re-ratings of MH's scoring of 27 candidates were not as off-target as standardization had suggested it might be. 5 were rated at the same score, 13 were one point out, 5 two points out, and 4 three points out, representing a correlation of 0.59. This is lower than the "rock-bottom" reliability co-efficient of 0.7 quoted by McNamara (2000. 58). Even if one point variance is regarded as being "acceptable", acceptability is still only 66% while 33% remain unacceptable. Measurement of grammatical accuracy was the key source of variance, with a correlation of 0.49, followed by manner (0.57), confirming difficulties envisaged from a theoretical perspective, medical knowledge (0.63), consultation pattern (0.65), and fluency (0.68). The suggestion is that the variation in performance level in this test population is not wide enough to warrant the number of levels, or breadth of the scale, adopted.

Nevertheless, in the words of Fulcher (1987. 291): "if a test is not reliable, it is not actually measuring anything, and so cannot possess validity". It must therefore be concluded that, on the evidence of sample re-ratings, the test lacks scorer reliability, and therefore it lacks validity. In addition, conclusions drawn from statistical analysis of scores, such as assessments of relative task difficulty, also lack validity.

A further element of reliability is "reliability of administration or implementation". As Weir (1993. 21) commented: "If the test is not well-administered, unreliable results may occur". On the day of the test, pairs of students were called to one of four testing rooms at pre-arranged times. It was thought to be agreed that once admitted to the testing room each candidate in each pair was given one doctor role card and one patient role card and allowed time to read the information provided while the previous pair perform their role play. In this way, it was supposed that time would not be lost while the candidates sat and read their role cards before beginning the interaction, a problem experienced in previous years.

This arrangement was, however, the cause of a serious problem affecting performance and threatening test validity. RE was unhappy with the idea of having the next pair in the test room while the previous pair was being tested, and apparently allowed the candidates to read their role cards outside the test room, allowing other candidates to view them. I began the session by insisting that the next pair not compare role cards or check their notes, but was finally successful only in the former, since cautioning of forthcoming candidates proved difficult without distracting the pair being tested. While MH prevented both practices from occurring, it seems that RE does not believe it makes any difference to performance. However, even if examiners feel the test is fair in this regard, the large number of complaints about these unreliable practices in student questionnaires show that face validity has been undermined.

Practicality

This kind of problem is related to practicality issues. Considering the "enormous practical constraints on the large-scale testing of spoken language proficiency" mentioned by Weir (1993. 40), special arrangements had to be made for scheduling. Testing 98 students (49 pairs), with ten minutes per pair, required 490 minutes of testing. The ninety minutes of class time available to each member of the team meant that only 36 pairs could be tested in the 8:45-10:15 am. session, with some pairs being tested in break time (10:15-10:30). The remaining13 pairs were tested in a later session by MH and TA in an afternoon session 4:15-5:30, RE and BZ being unavailable. In other words, the lack of human resources and time placed severe strain on the "test infrastructure", and this seemed to contribute to problems of reliability of implementation. It is also possible that it affected scorer reliability since, as MH points out, examiners needed more time to consider scores before testing the next pair.

[4. Recommendations and reflections]

There appear to be four main problems with the test and they concern scorer reliability, assessment criteria, reliability of implementation and practicality. I feel that rater standardization is the most serious of these, a fact admitted by MH. In contrast, RE believes that "the enormity of effort that would be required to standardize (training, cross-rating etc.) would not be so terribly productive". This view would likely find support among some commentators such as Matthews (1990. 120) who states that "the very attempt to measure proficiency by reference to behavioural criteria is basically flawed". However, her proposals for the abandonment of measurement of productive skills in favour of non-linguistic measurement of performance, such as successful diagnosis, would tend to shift the balance of the test towards emphasis on the medical component at the expense of speaking skills.

Nevertheless, even if we accept that some degree of inconsistency is unavoidable, it should, and could be reduced to a minimum through a couple of hours of training and discussion. As RE suggests, a return to a letter grade system (A, B, C, or D), or a reduced number of levels against which to measure performance might lead to improved scorer reliability. In terms of face validity, it would be assuring for the students to include a note in the next edition of the course book to the effect that rating had been standardized to allay fears and suspicions of rater inconsistency. As things stand these fears have been shown to be justified.

To begin with, as recommended by Weir (1993. 26), tape libraries should be established to "provide examples of performances at the prescribed levels". An initial rater standardization tape should begin with a recording of an example of a top-level performer, followed by a low-level (failing) performance to provide an outline of the expected range.

I also feel that the assessment criteria should be improved. This would have to begin with a reassessment of what is involved in the speaking skill as a consequence of discussions at standardization sessions. I would, for example, be in favour of replacing the grammatical accuracy analytical scale with "overall communicative effectiveness" according to the model adopted by McNamara (1996. 253), to cover measurement of discourse management, improvisational skills and pronunciation.

A further suggestion would be to reduce the amount of input in the course by introducing only six diseases to make available an extra session for testing so each examiner would examine only 6 pairs in one ninety minute session. In this way, there would be no need to have the next pair reading their role cards in the waiting zone. An interesting alternative would be to administer the same test two weeks running, allowing candidates to be tested by another examiner, with another partner, and a different disease thus allowing for "parallel forms" reliability (Weir et al 1999 A20) to reinforce scorer reliability and both construct and content validity, particularly with regard to format. Unfortunately, it may prove both unpopular and impractical. In future, it would also be helpful to collect data from real consultations to give both students and examiners a clearer idea of what doctors actually say in the target domain to inform both syllabus and assessment.

To sum up, empirical evidence appears to confirm weaknesses in the test construct *a priori*. These include doubts surrounding the way the patient's performance affects the doctor's, the measurement of improvisational skills and manner, and the need for standardization of scoring. Although it is a good test, unfortunately it remains invalid until proved valid.

Bibliography

Fulcher, G. 1987. "Tests of oral performance; the need for data-based criteria. ELTJ. 41 (4): 287-291.

Holst, M and R. Evans. 2000. Medical English. An ESP Application. JALT Hokkaido 2000 Proceedings. JALT Hokkaido. Sapporo.

Holst, M, C. Brown, C. Glick,, J. Tomei, R. Evans, E. Hollanders, G. Morris, M. Mino, and "T. A". 2001. English on Call. 3rd. Edition. Hokkaido University School of Medicine. Hokudai Insatsu. Sapporo, Japan.

Hughes, A. 1989. *Testing for Language Teachers*. Cambridge Handbooks for Language Teachers. Cambridge. Cambridge.

Matthews, M. 1990. The measurements of productivity skills: doubts concerning the assessment criteria of certain public examinations. ELTJ 44 (2) 117-121.

McNamara, T. F. 1996. *Measuring Second Language Performance*. Longman. London and New York.

McNamara, T. F. 2000. *Language Testing*. Oxford Introductions to Language Study Series. Oxford. Oxford.

Upshur, J and Turner C. 1995. Constructing rating scales for second language tests. ELTJ 49 (1): 3-12.

Weir, C. J 1993. *Understanding and Developing Language Tests*. London: Prentice Hall International.

Weir, C. J., Porter, D., Green R., 1999 *Language Testing*. Module Materials for MA TEFL School of Linguistics and Applied Language Studies. Reading, UK. The University of Reading.

White, R. 1999. The MA in TEFL by Distance Study Student Handbook, University of Reading. Reading.