ED 402 776 FL 024 334

AUTHOR Brown, James Dean; And Others

TITLE On JALT 95: Curriculum and Evaluation. Proceedings of the JALT International Conference on Language Teaching/Learning (22nd, Nagoya, Japan, November 1995). Section Seven: Testing and Evaluation.

PUB DATE Sep 96

NOTE 61p.; For complete document, see FL 024 327.

PUB TYPE Collected Works - Conference Proceedings (021) -- Tests/Evaluation Instruments (160)

LANGUAGE English; Japanese

EDRS PRICE MF01/PC03 Plus Postage.

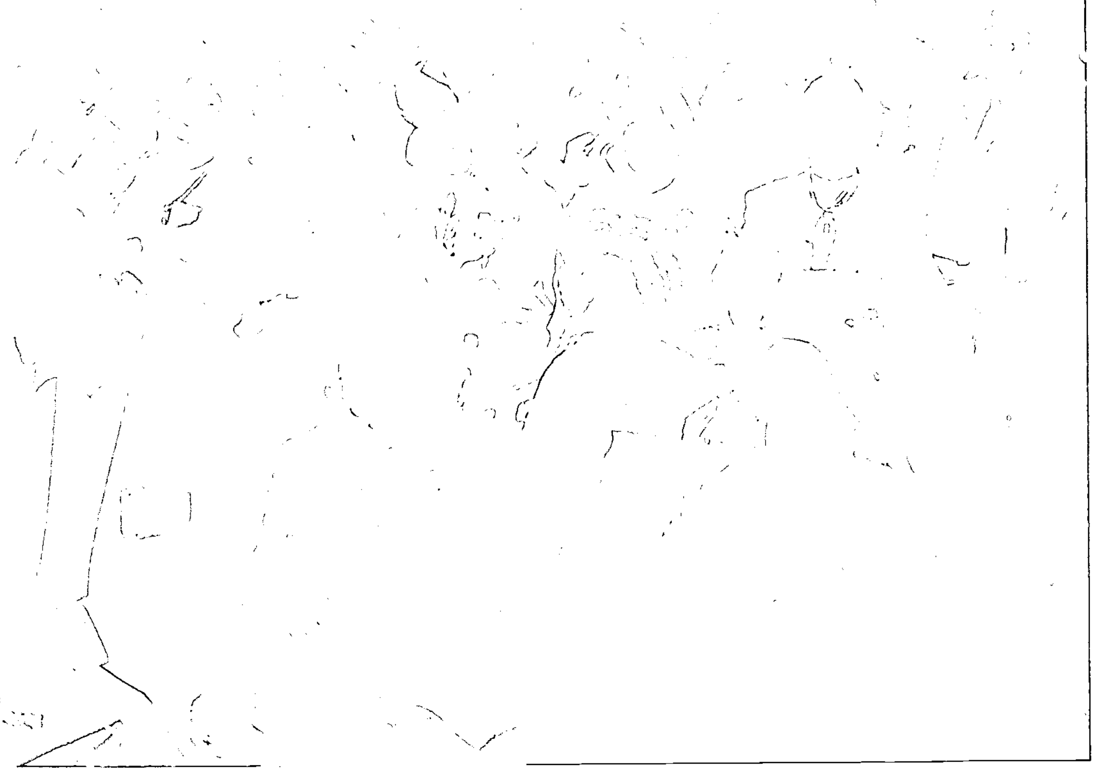
DESCRIPTORS Body Language; Cognitive Style; College Entrance Examinations; Communicative Competence (Languages); Course Evaluation; English (Second Language); Evaluation Criteria; Evaluation Methods; Foreign Countries; Higher Education; Instructional Effectiveness; Instructional Material Evaluation; *Language Tests; Learning Processes; Listening Skills; Nonverbal Communication; Questionnaires; Secondary Education; Second Language Instruction; *Second Languages; Simulation; *Student Evaluation; Student Evaluation of Teacher Performance; Task Analysis; *Testing; Test Reliability; Verbal Tests

ABSTRACT

This section contains conference papers on testing and evaluation in second language teaching. They include: "English Language Entrance Examinations in Japan: Problems and Solutions" (James Dean Brown); "Reliability and a Learner Style Questionnaire" (Dale T. Griffee); "Does It Work?" Evaluating Language Learning Tasks" (Rod Ellis); "Communicative Oral Testing" (Marion Delarche, Nicholas Marshall); "Evaluation of Gestures in Non-Verbal Communication" (Barry O'Sullivan); "Our Experiments in Oral Communication Tests" (Shuichi Yonezawa); "Simulations: A Tool for Testing 'Virtual Reality' in the Language Classroom" (Randall S. Davis); "Evaluation of Listening-Focused Classes" (Yoshinobu Niwa, Kazuo Iwata); and "Interpreting Teacher and Course Evaluations" (T. R. Honkomp). Individual papers contain references. (MSE)

Section Seven

Testing and Evaluation



English Language Entrance Examinations in Japan: Problems and Solutions

James Dean Brown
University of Hawaii at Manoa

For years, EFL teachers in Japan have recognized that many Japanese students study English for the primary, or even sole, purpose of passing high school or university entrance exams. Furthermore, most of the EFL teachers I have talked to about this issue say, in one way or another, that the English language exams have a negative effect on their teaching. In particular, many teachers say that both the content of the exams and the types of questions negatively impact their teaching and the language learning of their students. If this is a pervasive situation, and I think it is, then the EFL teachers in Japan should be in open rebellion. However, since open rebellion is not likely in this particular context, teachers should at least arm themselves (by learning as much as they can about the entrance examination system) so they can protect themselves and their students from the negative effects of the entrance exams on language teaching.

To that end, a Japanese colleague and I wrote two articles that:

1. described the 1993 entrance examinations at 21 universities including 10 public, 10 private, and the "Center" exam (Brown & Yamashita, 1995a), and
2. further investigated the 1994 exams at the same universities and how they differed from the 1993 exams (Brown & Yamashita, 1995b).

In other articles, I have:

3. argued for the use of listening tests on the university entrance exams (Brown & Christensen, 1987),
4. shown how test results are sometimes misinterpreted in Japan (Brown, 1993),
5. discussed the nature of *examination hell*, the social and psychological consequences of this exam system, the effects of entrance exams on adolescent life, the egalitarian roots of the exams, the relationship of the exams to career opportunities, the nature of *jukus* and *ronin*, the responsibilities involved in making decisions with such exams, and the washback effect of the English language entrance exams on EFL teaching (Brown, 1995a),
6. provided English definitions for some of the primary Japanese terminology that students use to describe *examination hell*, the examination system, and the examination preparation industry (Brown, In press),
7. discussed the *washback* effect of the university entrance exams on English language teaching in Japanese high schools (Brown & Kay, 1995), and
8. raised a number of these entrance examination issues in the public eye in Japan (Brown & Gorsuch, 1995).

But the purpose of my speech today is not to brag about all the publications I have written on the issue. Rather, I want to focus from a language testing perspective on some of the specific problems that the English language entrance exams have, and more importantly, I want to explore how these problems can be solved. Following the advice I gave in my own language testing book (Brown, 1995d), I will examine issues related to item quality, test revision strategies, test reliability, and test validity. I will also propose an agenda for change including discussion of openness issues, test development standards, professional development and scrutiny, and the need for much more research. I hope that discussion of these issues and any reform that results from such discussion will eventually help to put the university entrance examination "system" in Japan on a much more solid footing.

Item Quality

Item Quality Problems

In many entrance exam situations in Japan, a group of English teachers is given the task of producing a test that will be used for deciding who will be admitted to their university or deciding what level of English the students should study in that university. These are important decisions about the students' lives, yet these teams of test writers often have little or no experience in writing language tests, the test writers seldom receive guidance in how to write the items, and worse yet, the people are kept isolated from the rest of the world for security reasons.

In my experience, even professional test-item writers can only *estimate* the level and content of test questions that will be appropriate for a given group of students. As a result, even professional test-item writers will produce many items that are ineffective and do not work well with a particular group of students. In my experience, the number of ineffective items usually amounts to about one-third to one-half of those written. Since even professional item writers in the United States and elsewhere produce many items that are ineffective, I would assume that inexperienced item writers in Japan do so, too.

Item Quality Solutions

The solution to the problem of ineffective items is to pilot the test questions and perform item analysis on them. In fact, from a North American perspective, a test that remains unanalyzed is not worth giving to the students because, without item analysis, testers have no way of knowing how a set of items fits a particular group.

One problem that may occur, if items are not piloted, is that many of the items may be too difficult or too easy for the group of students being tested. Such items will not help in building a test at the appropriate level for spreading the students out into a normal distribution. A simple statistic called *item facility* (also known as item difficulty or item easiness) can be used to examine this issue and solve this potential problem.

Another problem that may occur, if items are not piloted, is that even those items at the right level of difficulty for the group may, for some reason, act quite differently from the rest of the items, that is, the low proficiency students may be answering them correctly, while the high proficiency students answer them incorrectly. A simple statistic called *item discrimination* can be used to examine this issue and solve this problem.

In short, in my view, failing to pilot the items used on entrance examinations borders on being unethical and is definitely unprofessional. After all, the entrance exams in Japan are used to make important decisions—decisions that will affect the children of Japan for the rest of their lives. Why is it, then, that the test designers cannot make the effort to make sure the test items they are using are of the best possible quality?

Test Revision

Test Revision Problems

From what many teachers have told me, the high school and university entrance examinations in Japan are seldom if ever revised or improved in any systematic manner. As described above, even the best entrance exams are often developed by a team of inexperienced test writers in the following five steps (see the second list below to understand why the numbering is out of sequence):

1. carefully develop the test,
6. administer the test ,
7. score the test ,
8. report the scores to the students, and
10. publish the test.

These five steps (numbered to match the list below) are fine as far as they go, but they leave out five other crucial steps that could be used to make the quality of the tests much better.

Typically in the United States, we use the same five steps in developing our tests, but we add some very important steps as shown in boldfaced type in the list of steps that follows (for more details on these steps, see Brown, 1995c, or 1995d):

1. carefully develop the test,
2. pilot the test,
3. analyze the results of the pilot administration statistically,
4. select those items that fit the group being tested and discriminate well,
5. **revise the test based on the statistical analyses,**
6. administer the test **under optimum conditions,**
7. score the test **as reliably as possible,**
8. report the scores to the students,
9. **analyze the final results statistically**, and
10. publish the test **and a technical manual that describes the test development, norms,**

reliability, validity, etc..

According to my information, the high schools and universities in Japan typically develop their entrance examinations using only steps one, part of six and seven, all of eight and part of 10, that is, the teachers on the testing team carefully develop the test; then they administer and score it and report the scores to the students; finally, they publish the test for public scrutiny (for examples, see Koko-Eigo Kenkyu, 1994a and 1994b).

These observations mean that the entrance examinations in Japan are most often not piloted, analyzed statistically, or revised. In addition, according to my information, the test administrations are often done under less than optimum conditions and the scoring is often less than maximally reliable. Furthermore, statistical analyses are seldom applied to the final results or reported publicly in a manual. From my perspective as an American language testing professional, I find the entrance exam development practices unethical and unprofessional. If I developed a test in this way in the United States, I would be attacked professionally and perhaps legally as well. And, I would deserve both.

From my perspective, the problem is that many or even most of the high school and university entrance examination development teams are skipping far too many steps. In particular, because they are skipping steps two through five and the last parts of steps six and seven, all of step nine and much of ten, they and the public have no way of knowing anything about how well their entrance examinations functioned or how accurate they were in making decisions based on the exams.

Test Revision Solutions

The solution to this problem seems clear: All ten of the steps listed above should be used in developing the entrance examinations in Japan at each and every institution that wants the privilege of doing entrance testing.

When I have suggested this solution in lectures throughout Japan, teachers have raised the specter of test security; "Oh so sorry, we cannot analyze and revise tests because of test security. Is very big problem in Japan." The speakers appear to believe that such a statement ends any need for further discussion of the issue. But to me, this is a classic straw man argument. Test security is not the issue; test security is a straw man. The inability to provide test security while doing a responsible job of testing is the real issue.

Organizations like Educational Testing Service manage to pilot test items in various ways without compromising test security, as do many other organizations both public and private in the United States. And, I firmly believe that anything American organizations can do, Japanese organizations can also do—probably much more effectively—once the Japanese decide to do it.

Several strategies can be used to securely pilot test items. In fact, three come immediately to mind: geographical distancing, temporal distancing, and interspersion of items on operational versions of the tests. *Geographical distancing* involves piloting test items in a place geographically distant from the cite where the exams will ultimately be given. For instance, a university in Kyushu might work out an agreement with a university in Hokkiado to pilot each others' items. The goal would be for each university to build a pool of items with known statistical characteristics that test writers could draw on in creating new tests. *Temporal distancing* involves piloting items over a long period of time, building up a large pool of items with known statistical characteristics, and using those items at a later date (in ways that are not predictable). *Interspersion of items* on operational tests involves putting some "experimental" items on every version of the test, year after year, and building a pool of items (with known statistical characteristics) that test writers could draw on. Sets of experimental items might even be different across the tests of a particular administration as long as 100 or so students (representative of the whole range of abilities in the student population) took each set of experimental items. Unlike the rest of the test, the experimental items would not have to be published after the tests were administered because they are *experimental* and because *they are not counted in the students' scores.*

This issue of piloting items in a secure manner is an important one. In fact, lack of piloting is the single issue that makes Japanese entrance exams most different from exams created by trained psychometricians elsewhere in the world. I might understand the lack of secure piloting if people were telling me that Japanese high schools and universities do not have the resources necessary to produce decent tests, or that they do not have staff with the know-how to produce effective tests. At least, such statements would be honest. But, I cannot believe that test security is an insurmountable issue which eliminates the possibility of piloting items before using them.

In short, in my view, the problem lies in the fact that many, if not most, of the universities and high schools that administer entrance examinations are simply too traditional or too understaffed or too under-financed or too lazy to do what is necessary to produce professional quality tests. And, to me, that attitude borders on being unethical and is definitely unprofessional. After all, the entrance exams are used to make important decisions—decisions that will affect the children of Japan for the rest of their lives.

Test Reliability

Reliability Problems

Test reliability can be defined as the degree to which a test is measuring consistently. Whenever we measure anything, we would like that measurement to be consistent. If the post office is measuring the weight of a package to determine how much postage you should pay and the clerk puts it on the scale twice, you would want the weight to come up exactly the same both times (or at least be very similar). If the package turned out to weigh 400 grams one time and 700 the next, you would complain. The problem that you would be complaining about is one of reliability. Such a scale would not seem to be measuring reliably.

In language testing, we also want our scales to be reliable, that is, we want to get the same (or very similar) scores for each student if we administer a test several times, or if we use several forms of the same test.

It is a fact that all measurements have errors. The question is not whether a measurement tool makes errors, but rather how much error a particular scale will produce. Such errors are also found on all language tests so it is not a question of whether errors are likely to occur, but rather how much error we can expect. On the TOEFL for instance, ETS (1995) reports that we can expect about plus or minus 15 point fluctuations in students' scores 68 percent of the time by chance alone. If we want to be 95 percent sure, we can expect fluctuations of 30 points (plus or minus). Thus ETS recognizes that there is error in their test scores and has done the analyses necessary to estimate how much effect that error is likely to have on decision making.

In several articles, Yamashita-san and I have suggested that the university entrance examinations in Japan may lack reliability. O'Sullivan (1995), in a letter to *The Language Teacher*, suggested that we had no evidence that the entrance examinations were unreliable, to which we answered:

> ...it is primarily the responsibility of the test developers (not the general public or the teaching profession or Brown and Yamashita) to provide evidence of the reliability and validity of the tests.

As the American Psychological Association (CDSEPT, 1985) put it, "Typically, test developers and publishers have primary responsibility for obtaining and reporting evidence concerning reliability and errors of measurement adequate for the intended uses" (p. 19). To my knowledge, little if any such evidence exists for the entrance examinations in Japan.

I have requested such information from a number of institutions and never gotten any. Since I suspected that such evidence might simply not exist, I also sought access to data in order to study these issues myself. In all cases, I have encountered resistance, secrecy, and a total lack of cooperation. Ladies and gentlemen, a black hole of information exists about these important examinations from which no light seems to escape. I, for one, can only conclude that problems *may* exist with the reliability of these tests. Naturally, I would welcome studies of these issues, recommend them as a solution to current shortcomings, and would myself happily participate.

Reliability Solutions

Reliability problems are not difficult to solve. Test developers can and should demonstrate the reliability of their test(s) using statistical techniques; they can also enhance test reliability, and strengthen their decision reliability.

Demonstrate reliability. How can the reliability of a language test be demonstrated? Actually, that is quite simple. Three strategies are commonly used to estimate the reliability of a test:

1. *Test-retest reliability* is an investigation of the consistency of a test over time. A test is administered on two different occasions to the same group of students and a correlation coefficient is calculated between the two sets of scores. A high correlation coefficient (one approaching 1.00) indicates a high degree of test-retest reliability.

2. *Equivalent forms reliability* is an investigation of the consistency of a test across forms. Two forms of a test are administered to the same group of students and a correlation coefficient is calculated between the two sets of scores. A high correlation coefficient (one approaching 1.00) indicates a high degree of

equivalent forms reliability.

3. *Internal consistency reliability* is an investigation of the consistency of a test across items. A single test is administered to a group of students on one occasion. Then, a formula (for instance, K-R20, K-R21, Cronbach alpha, etc.) is applied to the results of that administration and a reliability estimate is found. A high reliability estimate (one approaching 1.00) indicates a high degree of internal consistency reliability.

All three of these strategies can be used to statistically estimate the reliability of language tests, but the most commonly applied is the internal consistency strategy, probably because it is the easiest to deal with logistically: the test developer does not have to administer a test twice to the same group of students, or develop and administer two forms of the test. Instead, internal consistency reliability is based on a single administration of a single test.

The TOEFL, which is virtually the only English as a second language proficiency test that is widely used in the United States for university admissions decisions, has been repeatedly shown to be very reliable. For instance, ETS (1995) reports a respectable overall score reliability of .94, which can be interpreted as meaning that the TOEFL is 94 percent reliable and six percent unreliable. How many of the Japanese entrance examinations can report their reliability at all, much less a reliability that high?

Studying the reliability of a test is very very easy. I simply do not understand why Japanese high schools and universities are not studying these issues for their exams on a yearly basis. I'm sure that the educators in these institutions want post office scales to be reliable. Why don't they seem to care enough to insure that their entrance exams are equally reliable?

Enhance test reliability. Many factors may threaten the reliability of a test. Poorly written items, unclear test directions, and badly produced audio tapes are all potential problems with a test that can reduce its reliability. Other factors having to do with scoring like unreliable ratings (for writing samples, translations, interviews, etc.), mistakes in the answer key, and errors in adding scores for various subtests may also reduce the reliability of a test. Still other factors having to do with the students themselves (for example, fatigue, stress, emotional distress, lack of motivation, etc.) may reduce the reliability of the test.

In general, responsible test developers in the United States and elsewhere in the world do everything they can to eliminate or at least reduce the effects of such factors on the reliability of their tests. I suggest a number of strategies for doing so in my language testing book (Brown, 1995d). However, as I stated above, even the best tests have some unreliability. As a consequence, some energy must be put into studying the reliability of every exam in order to find out the degree to which efforts to enhance the reliability have been successful and in order to find new ways to enhance it.

Strengthen decision reliability. Even after studying the reliability of the entrance exams and enhancing the test reliability, test developers must also take into account reliability issues directly related to the decisions they are making with the test. In the case of entrance examinations, those decisions typically involve deciding which students should be admitted and which should be rejected from a given institution. Decision reliability is important because, as Brown and Yamashita (1995a, p. 26) put it: Perhaps the single most important fact about these very competitive entrance examinations is that the results are used to make decisions about students' lives—important decisions. As such, the examinations must be of the highest quality if they are to be fair to the students. Enhancing decision reliability is primarily a fairness issue, and it involves using the standard error of measurement to make responsible decisions.

The *standard error of measurement* is a statistic (calculated from the standard deviation of a test and a reliability estimate). The standard error of measurement describes the unreliable variance of a test in interpretable, test-score points. As such, the standard error of measurement can be used as a band of scores plus-or-minus around a cut-point that represents the band of unreliable decision making around that decision point (with certain degrees of probability). Once that band of unreliable decision making has been identified, administrators can seek additional information about the students who fall within that band, so that decision reliability will be enhanced.

For example, as mentioned above, the standard error of measurement on the TOEFL is about 15 points. At the University of Hawaii, we require a TOEFL score of 500 for students to be admitted. However, we recognize that unreliable variation in scores amounts to a 15 point band plus or minus around that cut-point of 500—a 15 point band where unreliable decisions are likely to be made. Hence, for students down as low as 485, as a matter of institutional policy, we consider additional information.

In short, in my view, any failures to check the reliability of the entrance exams, to enhance the test reliability of these tests, and to strengthen their decision reliability (using the standard error of measurement) border on being unethical and are definitely unprofessional. After all, the entrance exams are used to make important decisions—decisions that will affect the children of Japan for the rest of their lives.

Test Validity

Test Validity Problems

Test validity is the degree to which a test is measuring what it claims to be measuring. For instance, if a particular university creates an English entrance examination that is designed to test overall English language ability, then that is exactly what the exam should measure, and if it does so, the exam is said to be valid. Unfortunately, a number of teachers have raised questions about the validity of the entrance examinations saying that they use out of date testing methods and are mismatched with language teaching curriculum in Japan.

Out-of-date testing methods. Many of the entrance examinations include large numbers of multiple-choice grammar questions. In the view of many ESL/EFL teachers around the world, such discrete-point grammar questions are so unrelated to the current theories and practices of language teaching that serious questions arise as to the validity of the entrance exams (for more on these issues, see Brown & Yamashita, 1995a and c).

Other out-of-date item types include translation tasks, of which there are a large number on the university entrance exams. As far back as 1961, Robert Lado (1961, pp. 32-33) questioned the validity of translation tasks. In his own words:

> The ability to translate is a special skill. People who speak a foreign language well are not necessarily those who translate most effectively, although there is a correlation between knowledge of the foreign language and the capacity to translate. Some whose control of a foreign language is defective are nevertheless able to translate written material at considerable speed and reasonably well. ... Consequently, a translation test is not valid as a test of mastery of a foreign language.

Another way that entrance examinations are out-of-date is in the way they are administered. Consider the fact that, while computer labs abound in Japan, computerized testing, which is being developed on both large and small scales in the United States and elsewhere, has not even been considered in the university entrance exams of Japan (for more on uses of computers in language testing, see Brown, 1992).

In short, the abundance of out-of-date multiple-choice grammar items and translation items, as well as the pencil-and-paper delivery systems used on the entrance exams all pose potential threats to the validity of these exams.

Mismatches with curriculum. Even the reading portions of the exams, which are sometimes reasonably well-written, are often based on very difficult texts which are unlike the simplified texts that students are accustomed to in their English classes (also discussed in Brown & Yamashita, 1995a and c).

In addition, listening comprehension subtests are seldom found on the entrance exams (as discussed in Brown & Christensen, 1987), and speaking components are unheard of. This lack seems strange given the recent Monbusho revisions which added aural skills to the high school English language curricula. As explained in Brown and Yamashita (1995c, p. 98):

> A contradiction has also developed between what is included on these university entrance examinations and the Monbusho (1989) guidelines implemented in April 1993 for junior and senior high school English teaching. The guidelines advocate the addition of listening and/or speaking to the curriculum, but our analysis indicates that only six universities [out of 21] in 1993 and four [out of 21] in 1994 included even a listening component.

What does this contradiction mean? Put simply, if the proposed Monbusho curriculum reforms are theoretically sound and worthwhile and the high school and university entrance exams are not testing what is now included in the curriculum, then the entrance examinations lack validity.

Excuses. What some apologists for the entrance exams have said is that testing listening, extended writing, or speaking would be too expensive. I think that is nonsense. Very high fees are charged for the entrance examinations. For instance, a Japanese friend of mine just paid

40,000 yen to register her son to take a private university exam. And, tens of thousands of students take these exams (with most students failing, but paying for the privilege). Where do all those millions of yen go? And, why doesn't that money go into developing effective and valid communicative language tests? As I put it elsewhere, (Brown & Kay, 1995)

> ...what the universities are saying in effect is that Japanese young people are not important enough for the universities to find sufficient resources to test them properly—even though the universities charge the students very high fees for taking tests.

All in all, many reasons exist for doubting the validity of the entrance exams in Japan. And, as with reliability, the responsibility rests with the test developers (not the general public or the teaching profession or Brown and Yamashita) to demonstrate the validity of their tests. As the American Psychological Association put it (CDSEPT, 1985, p. 13), "evidence of validity should be presented for the major types of inferences for which the use of a test is recommended."

Test Validity Solutions.

Educational institutions in Japan can pursue three solutions to the validity problems: each institution that gives entrance exams should study and demonstrate the validity of their exams; the validity of existing tests should be enhanced; and the decision validity of the tests should be strengthened.

Demonstrate validity. How can the validity of a language test be demonstrated? As with reliability, it is actually quite simple. Three strategies are commonly used to study the validity of a test:

1. Content validity - This validity strategy involves demonstrating clearly that the content of the test matches the content of the curriculum or the domain being tested. This strategy frequently involves expert judgments about the degree of match between the test items and curriculum goals and objectives.

2. Construct validity - This approach to the study of validity usually involves setting up an experiment to demonstrate that the test does indeed test the psychological construct it claims to be testing. This strategy sometimes takes the form of a differential groups study or an intervention study (for a full explanation, see Brown, 1995d).

3. Criterion-related validity - This method of studying validity involves comparing test results with some well-respected independent measure of the same construct. Such a study is considered *concurrent* if the new test and the criterion measure are administered at the same time. The study is termed *predictive* if the new test is being studied to see how well it predicts some measure taken at a later time.

All three of these strategies are commonly used to study the validity of language tests. However, the strongest validity arguments are those based on two or even all three of these strategies.

The TOEFL, which is virtually the only English as a second language proficiency test that is used in the United States for university admissions decisions, has been repeatedly shown to be valid. For instance, ETS (1995) presents evidence for the content, criterion-related, *and* construct validity of the TOEFL.

How many of the Japanese universities have studied the validity of their entrance examinations? Yet, apparently, studying the validity of a test is relatively easy. I really do not understand why Japanese institutions are not studying these issues for their exams on a yearly basis. Don't they care?

Enhance test validity. As mentioned above, the TOEFL has been shown to be valid using a variety of validity strategies. For years, those arguments sufficed, but then public and professional criticism of the test began to surface, most of which boiled down to the fact that TOEFL was out-of-date in terms of validity. For instance, at this very conference, Savignon's keynote address pointed to the lack of social meaning in the TOEFL. Clearly then, even with ample evidence of validity in the test manuals, the TOEFL has come under attack for being out-of-step with developments in communicative language teaching.

Educational Testing Service has responded admirably to such complaints by developing the *Test of Written English* (TWE) and *Test of Spoken English* (TSE) programs, thereby including both extensive writing and speaking skills in the TOEFL suite of tests. In addition, ETS has worked hard on the TOEFL 2000 project, which is a major effort to completely revamp and update the TOEFL. How many Japanese institutions can say that they have writing and speaking components or that they have worked as hard as ETS to enhance the validity of their entrance exams?

In addition, in the United States and elsewhere, ideas about performance testing and other alternative methods of testing have been explored in recent years so that the validity of our exams can be enhanced (for more information, see the special alternative assessment issue of *TESOL Journal*, Vol. 5, No. 1). Are any such efforts being made in Japan? I think the answer is a resounding *NO*.

Strengthen decision validity. Even after studying the validity of the entrance exams and enhancing that validity, test developers must also take into account validity issues directly related to the decisions they are making with the test. In the case of entrance examinations, the decisions are typically made about who should be admitted and who should be rejected from a given institution. Carefully considering decision validity involves setting the cut-point (or acceptable standard) for passing the exam in a rational manner, and using multiple sources and types of information.

As for *standards setting*, a number of rational strategies can be used to set cut-points on a test. Three main categories of standards-setting methods are available to test developers and decision makers:

1. *State mastery methods* set standards in a dichotomous manner. Students are either considered to have the trait being measured or not have it. Many problems have been associated with this method.
2. *Test-centered continuum methods* rely on expert judgements of the test content to set standards.
3. *Student-centered continuum methods* focus on expert judgements of student performance to set standards.

Have any of these strategies been used in Japan, or do the test developers simply decide on the pass-fail score because it feels right? The question entrance exam developers need to address is: how are standards set for the cut-points used in deciding who will be admitted and who will not? (For much more on standards setting, see Brown, 1995d.)

As for *multiple sources and types of information*, according to Fujita (1991, p. 155), a majority of universities, particularly the elite universities, admit students solely on the basis of their entrance examination scores. In the United States, none of the major admissions tests (for instance, SAT, ACT, GRE, or TOEFL) are meant to be used as the sole criterion for admissions to any university. Indeed, the user's manuals for these tests all make a point of warning against the practice of using a single test score for this purpose, saying further, in one way or another, that the test scores should be used along with other types of information like previous grade point average, letters of recommendation, interviews, essays written by the students, other test scores, etc. Going even further, I argue in several places (Brown, 1987, 1995d) that most academic decisions should be made on the basis of multiple test scores (with various types of tests including proficiency, placement, diagnostic, and achievement) along with other types of information (like personal interviews, school records, feedback from professors, etc.).

In discussing the National Council on Education Reform (NCER) report (1985), Shimahara (1991, p. 133) says:

> In short, NCER [1985] has heightened an awareness of the need for alternative methods of recruiting applicants for employment in government and private industry: 'multidimensional and diversified' strategies to evaluate individual abilities throughout individual careers and strategies to improve what the Japanese often refer to as *gakureki shakai* a social structure that places excessive emphasis on one's specific school background as a criterion for employment and promotion.

As part of this process, perhaps the Japanese high schools and universities should develop multidimensional strategies for their admissions decisions.

In short, in my view, any failure to study the validity of the entrance exams, to enhance the validity of these tests, and to strengthen their decision validity (using rational standards-setting methods, and multiple sources and types of information) border on being unethical and are definitely unprofessional. After all, the entrance exams are used to make important decisions—decisions that will affect the children of Japan for the rest of their lives.

An Agenda for Change

So far, I have pointed to some major problems that the entrance exams in Japan have—problems involving item quality, test revision, test reliability, and test validity. I have also suggested solutions to each of these sets of problems. I would now like to briefly discuss four areas of general testing policy that could also be improved: openness issues, test development standards, professional development and scrutiny, and the need for much more research. In my opinion, improvement in these four areas would help to enhance the entire entrance examination decision-making process.

Openness Issues

As pointed out in Brown and Yamashita (1995a & c), many institutions openly provide their examinations for publication on a yearly basis. Such publication of tests is laudable and useful because it allows for public scrutiny. However, that is not enough. These institutions are also responsible for making sure that their tests are efficient, reliable, and valid. I have a number of reasons to believe that many of the examinations may be weak in all three areas. Yet, as I pointed out earlier, a black hole of information exists about these important examinations. Unfortunately, without information to the contrary, I can only conclude that problems *may* exist with the efficiency, reliability, and validity of these tests. Openness about these issues would not only allow the high schools and universities to defend the quality of their tests but also force those that are not already doing so to analyze the efficiency, reliability, and validity of their tests.

In countries other than Japan, test developers commonly and openly provide technical information about the quality of their tests as well as practical information to help test takers and score users interpret the norms, especially with regard to any particular student's scores. Such openness helps to avoid the appearance of being secretive, sneaky, and dishonest, and promotes open and honest communication between the test developers and the general public.

In the United States, a watch dog organization called *FairTest* serves as a kind of consumer advocate for test takers, making sure that openness and honesty are applied to any examinations that affect young Americans in important ways. Perhaps such an organization would be worthwhile and useful in Japan. I called *FairTest* just before leaving for Japan, and they indicated that they are very willing to send information that might help people in Japan establish a similar organization here. The purpose of such an organization might be to monitor testing practices in Japan and serve as an advocate that takes the point of view of the consumer, that is, such an organization would actually work for the fair treatment of the students who take entrance examinations, and in the process, monitor the efficiency, reliability, and validity of the exams.

For anyone who is interested in contacting them, *FairTest's* phone number is 1-617-864-4810, their e-mail address is <fairtest@aol.com>, and their snail-mail address is:

FairTest
National Center for Fair & Open Testing
342 Broadway
Cambridge, MA 02139 USA

Professional Development and Scrutiny

Unfortunately, as I mentioned above, many or even most of the high school and university entrance examinations are developed by amateurs who know very little about this very specialized area called test development. Is it any wonder, then, that they do not know how to do a truly professional job of test development? Two general steps could be taken to help make such test developers more professional: first, establish standards for testing and, second, establish a systematic test review process.

Establish Standards for Testing. Many of the problems discussed in this speech are avoided by test developers in the United States because, as a profession, they follow the *Standards for Educational and Psychological Testing* (CDSEPT, 1985). This document, which clearly lays out the responsibilities of test developers, was developed jointly by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME).

Other such documents have been prepared independently by various organizations, for example, the Joint Committee on Testing Practices (1988), the Association for Assessment in Counseling (1993), and the National Council on Measurement in Education (1995) have all published their own guidelines for test developers and users, and the American Psychological Association (1986) has even published guidelines for computer-based tests.

Obviously, professional standards for test development and use are very important in the United States. The standards provided in these various documents help test developers to know what is expected of a good test and of them as test developers. Thus, test writers can do a better job of developing their tests. In short, the standards provided by various professional associations help American test developers to practice their trade in a professional manner. If it does not already exist, a similar set of standards fitting the conditions in Japan should immediately be developed by a consortium of important Japanese educational organizations.

Establish a Systematic Test Review Process. In addition, I have always felt that letting the high schools and universities in Japan monitor the quality of their own exams is roughly equivalent to letting the fox guard the chicken coop. In the United States, *Buros Mental Measurements Yearbook* (for example, Kramer & Conoley, 1992) is a periodic publication that provides a collection of reviews of published tests. *Buros* serves as a critical watch dog on all published tests used in North America. The effect of *Buros* reviews is to force openness and foster critical thinking about the tests that are developed for use in the United States. Is there such a regular publication in Japan?

I believe that both the *Standards* and *Buros* tend to keep test developers honest and professional. Similar institutions in Japan might have the same effects. The point is that the entrance examinations in Japan are far too important to be left entirely up to the test designers. Teachers and professors are not infallible; they must be held accountable, perhaps for the first time in history, for the important admissions decisions that they are making because those decisions are so profoundly important to young Japanese lives.

The Need for Much More Research

In his response to the Brown and Yamashita (1995a) article, O'Sullivan (1995, p. 256) suggested that further research should be done on the following three research questions:

1. Is there evidence of a topic awareness bias in some tests?
2. How harmful is the dependence on translation?
3. Can we establish the content and construct validity of these tests?

While interesting, his questions seem a bit too specific and narrow for the immediate research needs vis-a-vis the entrance examinations in Japan. The following research questions are liberally adapted and expanded from Brown and Yamashita (1995d). I hope that they will form at least a start on a research agenda for studying the entrance examinations in Japan:

1. How well do the items on the entrance examinations perform in terms of item facility and discrimination? What statistics should be used to help in selecting items for the entrance examinations? What types of items should be used to improve the quality of the tests and make them more valid?
2. What test development and revision practices are followed in creating the entrance examinations? Would the exams be improved by following the ten steps listed in this speech? What would be the effects on reliability and validity of such revision processes?
3. How are norms established on these tests, and how do they vary from institution to institution and year to year?
4. What evidence exists for the reliability of these entrance examinations (for instance, what is the K-R20, or Cronbach alpha reliability of these tests)?
5. What evidence is there for the decision reliability of these exams (that is, what is the standard error of measurement, and how is it used, if at all, to make admissions decisions responsible and fair, and are additional types of information used for students who fall within this band of unreliable test score variance)?
6. What evidence is there for the content, construct, criterion-related, face, decision, or social validity of these tests (for more on these types of validity, see Brown, 1995b or 1995c)?
7. What evidence is there for the decision validity of the entrance examinations? How are standards set for the cut-points used in deciding who will be admitted and who will not? Are state mastery methods used? Or, test-centered continuum methods? Or, student-centered continuum methods? Are rational methods used at all? (for more on standards setting, see Brown, 1995d) Are multiple sources and types of information used to strengthen the decision validity of the entrance examinations?
8. Why do the examinations cost so much given the relatively cheap and easy-to-score formats that are used? Or put another way, why is it that communicative listening and speaking subtests are not used on these exams even though there is apparently plenty of revenue to support such sound testing practices?
9. What is the impact of the *washback* effect of these tests on the educational system in Japan? In particular, what is their effect on

the teaching of English?

If you already have answers to all of these questions about the entrance exams in Japan, then I apologize; you are doing a fine job. But, if you do not have answers to all of them, it is time to get to work. Failure to do so would be irresponsible.

In fact, in my view, any failure to pilot, analyze, and revise the entrance exams, any failure to check and enhance the reliability of these tests, or failure to strengthen the decision reliability of the tests, any failure to verify and enhance the validity of the exams, or failure to study the decision validity of the exams, any failure to be open, to development testing standards, to insure professional development and scrutiny, or to do the much needed research, any such failures border on being unethical and are definitely unprofessional. After all, the entrance exams in Japan are used to make crucially important decisions—decisions that will affect the children of Japan for the rest of their lives.

References

American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC: American Psychological Association.

Association for Assessment in Counseling. (1993). *Multicultural assessment standards: A compilation for counselors.* Alexandria, VA: American Counseling Association.

Brown, J. D. (1987). False beginners and false-starters: How can we identify them? *The Language Teacher, 11*(14), 9-11.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design.* London: Cambridge University Press.

Brown, J. D. (1990). Where do tests fit into language programs? *JALT Journal, 12*(1), 121-140.

Brown, J. D. (1992). Using computers in language testing. *Cross Currents, 19*(1), 92-99.

Brown, J. D. (1993). Language testing hysteria in Japan. *The Language Teacher, 17*(12), 41-43.

Brown, J. D. (1995a). English language entrance examinations in Japan: Myths and facts. *The Language Teacher, 19*(10), 21-26.

Brown, J. D. (1995b). Differences between norm-referenced and criterion-referenced tests? In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 12-19). Tokyo: Japan Association for Language Teaching.

Brown, J. D. (1995c). Developing norm-referenced language tests for program-level decision making. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 40-47). Tokyo: The Japan Association for Language Teaching.

Brown, J. D. (1995d). *Testing in language programs.* New York: Prentice-Hall.

Brown, J. D. (1995e). *The elements of language curriculum: A systematic approach to program development.* New York: Heinle & Heinle Publishers.

Brown, J. D. (In press). A *gaijin*'s guide to the vocabulary of entrance exams. *The Language Teacher.*

Brown, J. D., & Christensen, T. (1987). Interview: James D. Brown. *The Language Teacher, 11*(7), 6-10.

Brown, J. D., & Gorsuch, G. (1995). An interview with J.D. Brown: Analyzing the value, meaning of entrance examinations. *The Daily Yomiuri,* No. 16271, October 30, 1995, p. 15.

Brown, J. D., & Kay, G. (1995). English language entrance examinations for Japanese universities: Interview with James Dean Brown. *The Language Teacher (JALT), 19*(11), 7-11.

Brown, J. D., & Yamashita, S. O. (1995a). English language entrance examinations at Japanese universities: What do we know about them? *JALT Journal, 17*(1), 7-30.

Brown, J. D., & Yamashita, S. O. (1995b). *Language testing in Japan.* Tokyo: Japan Association of Language Teaching.

Brown, J. D., & Yamashita, S. O. (1995c). English language entrance examinations at Japanese universities: 1993 and 1994. In J. D. Brown & S. O. Yamashita, *Language testing in Japan* (pp. 86-100). Tokyo: Japan Association of Language Teaching.

Brown, J. D., & S. O. Yamashita. (1995d). The authors respond to O'Sullivan's letter to *JALT Journal*: Out of criticism comes knowledge. *JALT Journal, 17*(2), 257-260.

CDSEPT (Committee to Develop Standards for Educational and Psychological Testing). (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

ETS (Educational Testing Service). (1995). *TOEFL test & score manual: 1995-1996 edition.* Princeton, NJ: Educational Testing Service.

Fujita, H. (1991). Education policy dilemmas as historic constructions. In B. Finkelstein, A. E. Imamura, & J. J. Tobin (Eds.), *Transcending stereotypes: Discovering Japanese culture and education* (pp. 147-161). Yarmouth, ME: Intercultural Press.

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education.* Washington, DC: American Psychological Association.

Koko-Eigo Kenkyu. (1994a). *'94 Kokukoritsu-dai:Eigomondai no tetteiteki kenkyuu.* Tokyo: Kenkyusha.

Koko-Eigo Kenkyu. (1994b). *'94 Shiritsu-dai: Eigomondai-no tetteiteki kenkyuu.* Tokyo: Kenkyu-Sha.

Kramer, J. J., & Conoley, J. C. (Eds.), (1995). *The twelfth mental measurements yearbook.* The Buros Institute of Mental Measurements, Lincoln, NE: University of Nebraska.

Lado, R. (1961). *Language testing: The construction anduse of foreign language tests.* New York, NY: McGraw-Hill.

NCER (National Council on Education Reform). (1985). *Report.* In Gyosei (Ed.) *Rinkyoshin to kyÇiku-kaikakujiyuka kara koseishugi e* (From liberalization to putting an emphasis on individuality). Tokyo: Gyosei.

National Council on Measurement in Education. (1995). *Code of professional responsibilities in educational assessment.* Washington, DC: National Council on Measurement in Education.

O'Sullivan, B. (1995). Reaction to Brown and Yamashita. *JALT Journal 17*(1), 255-257.

Shimahara, N. (1991). Examination rituals and group life. In B. Finkelstein, A. E. Imamura, & J. J. Tobin (Eds.) *Transcending stereotypes: Discovering Japanese culture and education* (pp. 126-134). Yarmouth, ME: Intercultural Press.

Reliability and a Learner Style Questionnaire

Dale T. Griffee
Seigakuin University

Recently, interest in classroom research has been on the rise and many classroom researchers are calling for the reliability reports of research instruments such as achievement tests, interviews, questionnaires, and surveys (Chaudron, 1988; Hatch & Lazaraton,1991; Kasper & Dahl, 1991). Nevertheless, despite these calls, it is not yet common for classroom researchers to include the reliability figures of their research instruments resulting in methodologically flawed research (Chaudron, 1988; Long, 1990). As more classroom teachers engage in research, the issue of determining and reporting reliability will become more important. The purpose of this paper is to explain what reliability is, to illustrate how to determine reliability using an example of a Learning Style Questionnaire (LSQ) from Hinkelman & Pysock (1992), and using the same instrument, to illustrate how the reliability of a research instrument can be improved through instrument revision.

What is Reliability?

Reliability is a statistical procedure used to determine how consistent an instrument is. For the purposes of this paper the term "instrument" will be used to cover any means used by a teacher to elicit and gather data including achievement tests, questionnaires, surveys, and even interviews. If we look at various definitions of reliability given by researchers, the word that appears in almost every definition is the word "consistent" or "consistency" (Davies, 1990; Hatch & Farhady, 1982; Hatch & Lazaraton, 1991; Henning, 1987; Johnson, 1992; Oller, 1979; Seliger & Shohamy, 1989; Vierra & Pollock, 1992; Weir, 1990). The question that a reliability estimate seeks to answer is how consistent is this instrument? (Hatch & Lazaraton, 1991; Johnson, 1992; Seliger & Shohamy, 1989).

Reliability can be seen as a ratio between the true score and the error score (Bachman, 1990; Brown, 1995; Henning,1987; Hatch & Lazaraton,1991. A true score is what Brown (1995) calls "meaningful variance" by which Brown mean how much the student knows. An error score is what Brown calls "measurement error" which indicates how much error is in the test. Measurement error is produced by anything other than meaningful variance, such as the effect of the student's physical condition, the student's emotional condition, and the test-taking environment (e.g., how hot the room was on the day of the test). Measurement error also results from ambiguous questions, idiomatic language which may not be known or understood by the test takers, and difficult to understand instructions. In other words, when we look at the results of instruments such as tests, surveys, questionnaires, or even the ratings of student interviews, we should think of the score as representing what the student really knows (the true score) plus all the other factors that might interfere (the error score). Looked at in this way, reliability is the ratio of the true score (or meaningful variance) to error score (or measurement error).

Reliability can also be seen as a correlation between two sets of numbers (Davies, 1990; Henning, 1987; Hughes, 1989). As an example, suppose we have the scores for a listening test from a certain class. The test papers are accidentally thrown into the trash and we, with apologies to our students, administer the same test the following week. Then, to our surprise, the original test papers show up. Now we have the first test scores and another set of test scores, all from the same test, the same students, and only a week apart. The scores should be the same, but as we start looking we notice that many students received scores on the second test a few points higher and in some cases lower than the first test. We suppose that the difference is measurement error. We then line up the scores from the two tests to see exactly how they match. We can see some difference, but we wonder exactly how much difference there is. We enter the scores in a computer statistical program and push the correlation key and out comes a number. That number is a correlation coefficient which can range from minus one to plus and the closer it is to plus one, the better.

How is Reliability Related to Validity?

To be valid, a test must be reliable. You recall the listening test mentioned above in the discussion on reliability as correlation. My claim was that my test was a test of listening. In support of that claim, suppose that I gave reasons why my test is a listening test and not some other kind of test, for example a grammar test. What I am doing is making a claim for test validity. Validity has to do with the match between the stated purpose of a test and the actual function of the test, what the test actually tests. In other words, validity is an argument whereas reliability is a number. Validity is a claim and reliability is an indication of how adequately we are fulfilling the claim (Davies, 1990, p. 53). What a test is supposed to do is, according to Oller (1979, p. 4), also a question of validity, prompting Oller to conclude that validity can never exceed reliability. The relationship between reliability and validity is such that a research instrument can have test reliability without test validity, but it can never have test validity without test reliability (Weir, 1990, p. 33).

Types of Reliability

What types of reliability are there, when do we use which type, and how do we calculate the different types?

There are three types of reliability generally reported by researchers (Weir, 1990, p. 32). They are inter-rater reliability, internal consistency reliability, and parallel-forms reliability.

Inter-rater reliability is the measure of agreement among human test raters. Raters score the test (typically an interview or a composition) and their scores are correlated and the resulting correlation coefficient is taken as the reliability coefficient. Internal consistency reliability, on the other hand, uses statistics from the test such as the mean and standard deviation to calculate a reliability coefficient. The most common ways of calculating internal consistency reliability are the Kuder-Richardson formula 20 and Kuder-Richardson formula 21 and Cronbach's alpha formula. Parallel-forms reliability requires form A of a test and form B. While both forms must be different, they must be parallel or equivalent in every way. As a pretest at the beginning of the semester, half of your class receives form A and the other half receives form B. At the end of the semester as a final exam, your class takes the same test, but this time those who took form A are given form B. The two test forms are scored and the scores are correlated.

We now know how many types of reliability there are, but we do not know when to use which type. Seliger and Shohamy (1989, p. 185) say that which type of reliability to report depends on the type of data you are collecting. If you are collecting data which requires judgment calls such as an interview, the appropriate type of reliability to report would be inter-rater reliability. If you are using two forms of the same data collection instrument and you want to know if the forms are really equal, report parallel-forms reliability. If you are using an instrument which has many independent items and you want to know if all the items elicit the same information as would be the case if you were administering an achievement test or a questionnaire, report internal consistency reliability.

What is an Acceptable Reliability Coefficient?

For most educational research, Vierra & Pollock (1992, p. 62) say that .90 or better is very good, between .80 and .90 is acceptable, below .80 may be acceptable when the variable is known to be difficult to measure, and below .60 is not adequate. For inter-rater reliability, Allwright & Bailey (1991, p. 46) indicate that classroom researchers should strive for at least an .85 coefficient. A paper and pencil achievement test should be at least .90 (Davies, 1990, p. 22), but Reid (1990, p. 326) would allow a .70 coefficient for a difficult to measure trait such as learning styles.

Table 1
Acceptable reliability coefficients

Coefficient	Status
.90+	very good
.90 to .80	good
.80 to .70	acceptable if trait is difficult to measure
.70 or below	not acceptable

Why is it Necessary to Report Reliability?

Chaudron (1988) has stated that if classroom researchers create a research instrument to collect data, the first thing they have to demonstrate is the reliability of the categories they propose. He noted that researchers "have infrequently confirmed the reliability and validity of their observational measures" (1988, p. 23). Since then, Long (1990; p. 163) has echoed the call by noting that many second language acquisition studies are methodologically flawed by the lack of reliability data. There are at least four answers to the question, why is it necessary to report reliability? They are trustworthiness, generalizability, fairness, and revision.

1. The issue of trustworthiness is the degree of confidence one can have in the research (Henning, 1987, p. 74). There is simply no point in giving us results that we cannot trust (Davies, 1990, p. 23; Hatch & Lazaraton (1991, p. 529).
2. Generalizability is the degree to which we can use the results of research in situations other than the one in which it was performed. If findings are not reliable, there is no point in using them in other settings (Allwright & Bailey, 1991, p. 49).
3. Fairness. Many decisions affect the lives of students from passing or failing a course to who will be selected for an overseas study program. The more important the decision, the greater the reliability that must be demanded.
4. For a classroom researcher, instrument revision is one of the most important uses of reliability. Simply put, a low reliability coefficient indicates some sort of problem with the instrument (see Bachman, 1990, p. 160; Oller, 1979). A reliability study will not tell you what the problem is nor will it tell you how to solve the problem, but a low reliability coefficient acts as a red flag indicating danger.

To illustrate how reliability can be used to revise and improve an instrument, this paper now reports two studies dealing with the reliability of a learning style questionnaire on learner modalities. The statistics used to analyze the questionnaire are described and results are given which show low reliability especially in one portion of the questionnaire. The revision process is then described and the results of two follow-up studies are given. It is concluded that reliability is not only a necessary statistic to report, but helpful to the revision process.

Results of the first study

Thirty-three second-year university students participated in the first pilot study which was administered to 16 males and 17 females. Two students, one male and one female dropped out leaving a total of thirty-one students in the study.

The Learning Style Questionnaire (LSQ) from Hinkelman and Pysock, (1992) is titled, What is your learning style? and can be found in Appendix 1. Students were asked to complete each of twelve sentences by awarding a score of 3 points to the best answer, 2 points to their second best answer and 1 point to the least preferred answer. Scores can range from a minimum of twelve to a maximum of thirty-six points. The score at the bottom of the first column indicates the degree of preference for the visual modality, the score at the bottom of the middle column indicates the preference for the auditory modality and the score for the third column indicates the preference for the kinesthetic modality.

To estimate reliability, Cronbach alpha, a split-half procedure which measures internal consistency, was chosen because it is effective for weighted scores. The assumption of Cronbach alpha is normal distribution. For the formula and discussion of this statistic, see Brown (1995). The formula was manually put into a spread sheet computer program. Once the formula was verified using figures provided in Brown (1995), new data could be entered and the formula recalculated.

The reliability coefficients are given in Table 2 in terms of the visual (V), the auditory (A), and the kinesthetic (K) sections of the questionnaire. Since learning modalities are difficult to measure, a .76 reliability coefficient can be considered acceptable for the visual and kinesthetic sections, but the reliability coefficient of the auditory section is clearly inadequate.

Table 2
Reliability coefficients

	V	A	K
Cronbach a	.76	.40	.76

The Revision Process

Three strategies were identified which would increase the reliability of the LSQ instrument (Reid, 1990). These strategies are to increase item homogeneity, to increase the number of items, and to pair and correlate items.

Item Homogeneity. The key idea behind this strategy is that the more similar the items types are, the higher the reliability (Henning, 1987; Davies, 1990). Rewriting the items to make them more alike (homogeneous), makes them easier to understand which will, in turn, lead students to answer them in a more consistent way. In practice, item homogeneity means simplifying stems ("When I am bored, I . . . " was revised to "In class I sometimes . . .") and eliminating multiple examples ("In my free time I like to read, draw, watch TV" was revised to "In my free time I like to read").

Increasing the Number of Items. Increasing the number of items gives a wider range of scores which will increase reliability. More items will also give the researcher a chance to eliminate those items not working well and still leave enough items that are working well.The first LSQ instrument (Appendix 1) had 12 items. This was increased to 36 items in the second version (see Appendix 2).

Item Pairing. All items for the second version were written in pairs and then randomly placed. Thus, item 14a is paired with item item 33a and item 8a is paired with item 13a as shown in table 3.

Table 3.
Examples of revised paired items

Item			
14	I learn best		
	a_____in the library	b_____in the language lab	c_____outside
33	I learn best		
	a_____in class reading/writing	b_____in class discussions	c_____in class projects
8	I like		
	a_____watching animals	b_____listening to animals	c_____touching animals
13	At the zoo, I like		
	a_____looking at the	b_____hearing the animals sounds	c_____petting the animls

Results of the Second Study

A total of thirty-three students (19 men and 14 women) participated in the second pilot study. The LSQ instrument was the revised version two of the previous instrument (see Appendix 2). The revised instrument was thirty-six questions long and it was administered and scored in the same way as in the first pilot.

Cronbach alpha was used to determine reliability and the Pearson product moment correlation formula was used to correlate randomly paired items. The correlation results are shown in Appendix 4. The reliability coefficients results are given in Table 4. While the auditory coefficient is still lower than the visual and kinesthetic coefficients, version two can be considered an adequately reliable instrument although at 36 questions long, it may not be as convenient for classroom use as the shorter version.

Table 4
Reliability coefficients results for Version Two

	V	A	K
Cronbach a	.91	.76	.89

Discussion

Pairing and then correlating the paired items make it possible to identify which items to retain, which items to revise, and which items to eliminate. All item pairs within each area of the LSQ were correlated using the Pearson formula (StatView 4.2 for the Macintosh). Specifically, all items pairs within the visual section were correlated, all item pairs within the auditory section were correlated, and all item pairs within the kinesthetic section were correlated.

An item was eliminated if two of the three possible correlations were not statistically significant at p <.05. For example, (see Table 3) item 14a was correlated with item 33a, item 14b was correlated with item 33b, and item 14c was correlated with item 33c. The results of that correlation are listed in Appendix 4. Looking again at Table 3, the correlations of the visual, the auditory, and the kinesthetic parts of items 14 and 33 are not statistically significant. and both these items were rejected. The correlations of items 8 and 13, on the other hand, are statistically significant and were included in revised versions of the LSQ instrument. Using this criteria, twelve pairs were eliminated leaving six pairs or twelve questions in version three (see Appendix 3). The resulting reliability coefficients for version three were recalculated using the Cronbach alpha formula. The results were .86 for the visual section, .75 for the auditory section, and .86 for the kinesthetic section. These correlations can be taken as reliability coefficients and indicate that either the long version of the LSQ with 36 questions (version two in Appendix 2) or the short version with 12 questions (version three in Appendix 3) may confidently be used with student populations that are similar to the students described in this study.

Conclusion

This paper has shown the important role reliability can play in instrument revision. Revision is important because teacher-researchers create data elicitation instruments based on the best knowledge available to them at the time. The studies reviewed in this paper clearly show, however, that teacher-researcher intuition while necessary, is not sufficient. Teacher-researchers require feedback to guide the revision and improvement of their data elicitation forms. Reliability studies can provide the basis for that feedback.

References

Allwright, D. & Bailey, K. M. (1991). *Focus on the language classroom*. Cambridge: Cambridge University Press

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. Cambridge: Cambridge University Press.

Brown, J. D. (1995). *Testing in language programs*. Englewood Cliffs, NJ: Prentice-Hall.

Chaudron, C. (1988). *Second language classrooms: Research on teaching and learning*. Cambridge: Cambridge University Press.

Davies, A. (1990). *Principles of language testing*. Oxford: Basil Blackwell.

Hatch, E. & Farhady, H. (1982) *Research design and statistics for applied linguistics*. Cambridge: Newbury House.

Hatch, E. & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Newbury House.

Henning, G. (1987). *A guide to language testing*. Boston: Heinle & Heinle.

Hinkelman, D. & Pysock, J. (1992). The need for multimedia ESL teaching methods: A psychological investigation in learning styles. *Cross Currents 19* (1), 25-35.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Johnson, D. M. (1992). *Approaches to research in second language learning*. New York: Longman.

Kasper, G. & Dahl, M. (1991). Research methods in interlanguage pragmatics. *Studies in second language acquisition 13* (2), 215-247.

Long, M. (1990). Second language classroom research and teacher education. In C. Brumfit, & R. Mitchell, (Eds.), *Research in the language classroom* (pp 161). ELT documents 133. Modern English Publications in association with The British Council.

Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.

Oller, J. W. Jr. (1979). *Language tests at school*. London: Longman.

Reid, J. (1990). The dirty laundry of ESL survey research. *TESOL Quarterly 24* (2), 323-338.

Richards, J., Platt, J.. & Platt, H. (1992). *Longman dictionary of language teaching and applied linguistics* (2nd Ed.) Harlow: Longman.

Seliger, H. W. & Shohamy, E. (1989). *Second language research methods*. Oxford: Oxford University Press.

Shohamy, E. (1994). The role of language tests in the construction and validation of second-language acquisition theories. In E. Tarone, S. Gass, & A. Cohen (Eds), *Research methodology in second-language acquisition* (pp. 133-142). Hillsdale, NJ: Lawrence Erlbaum Associates.

Vierra, A. & Pollock, J. (1992). *Reading educational research*. Scottsdale: Gorsuch Scarisbrick.

Weir, C. J. (1990). *Communicative language testing*. London: Prentice Hall.

Appendix 1

What is your learning style? v1

Name _______________ Student Number _______________

There are 3 answers in each line. Write number "3" next to the answer you like best. Write number "2" next to the answer you like second best and write number "1" next to the answer you like third best.

1. I learn best by
a._____seeing something written b._____listening c._____doing it myself

2. To find a place, I want someone to
a._____draw me a map b._____tell me in words c._____take me there

3. In my free time, I like to
a._____read, draw, watch TV b._____talk, listen to music c._____play sports, drive, cook

4. I make a plan by
a._____writing notes b._____talking to others c._____just do it

5. When I want to talk to my friends, I like to
a._____write them a letter b._____telephone c._____visit them

6. I am good at
a._____drawing or math b._____talking with people with machines c._____working

7. I like to learn a computer by
a._____reading the manual b._____having a teacher explain it c._____doing it myself

8. After a good party, I want to
a._____look at photos of the party b._____talk about the party c._____have another party

9. I like
a._____color and design b._____music, bird sounds c._____moods, feelings

10. In any class, I like to
a._____use the textbook b._____listen to the teacher c._____do exercises

11. I like to learn English by
a._____watch English videos b._____listening to a tape c._____talking to a native speaker

12. In general, I am
a._____a watcher b._____a listener c._____a doer

TOTAL
a.__________ b.__________ c.__________

Appendix 2

What is your learning style? version two

Name ________________________ Student Number________________________

There are 3 answers in each line. Write number "3" next to the answer you like best. Write number "2" next to the answer you like second best and write number "1" next to the answer you like third best.

1. I like to learn English by
a._____reading the textbook b._____talking in pairs c._____moving my body

2. I enjoy
a._____drawing b._____singing c._____dancing

3. When I am lost, I like to
a._____look at a map b._____ask someone directions c._____go with someone

4. I like to
a._____see the words b._____say the words c._____move my hands with the words

5. I like
a._____using textbooks b._____listening to tapes c._____doing dramas

6. I learn best when I
a._____see something b._____hear something c._____touch something

7. In my free time, I like to
a._____see videos b._____phone my friends c._____play sports

8. I like
a._____watching animals b._____listening to animals c._____touching animals

9. I plan something by
a._____making a list b._____discussing it c._____practicing it

10. In class at school, I like
a._____demonstrations b._____explanations c._____practice exercises

11. I learn best when I
a._____look at something b._____say something c._____touch something

12. To find a new place, I say
a._____"draw a map for me" b._____"tell me the way" c._____"take me there"

13. At the zoo, I like
a._____looking at the animals b._____hearing the animal sounds c._____petting the animals

14. I learn best
a._____in the library b._____in the language lab c._____outside

15. I like teachers who
a._____write clearly on the board b._____speak clearly c._____give worksheets to write on

16. I learn best by
a._____reading words b._____hearing words c._____acting with words

17. I like to
a._____write to friends b._____telephone friends c._____travel to friends

18. I like to contact friends by
a._____writing a letter b._____calling on the telephone c._____going to their house

19. To learn a computer, I first
a._____read a book about it b._____listen to someone c._____touch the keys

20. At a party I want to

a._____take pictures b._____sing songs c._____play games

21. I learn best by
a._____reading stories b._____hearing stories c._____acting stories

22. When I cook, first I usually
a._____look at a cookbook b._____have someone tell me c._____pick up the food

23. I like to
a._____look at the board b._____listen to the teacher c._____stand up and practice

24. I like teachers, who
a._____use pictures b._____let us discuss c._____make us move around

25. At a party, I want to
a._____look at photos b._____hear people tell stories c._____eat snacks

26. I make a plan by
a._____writing notes b._____listening to others c._____walking and thinking

27. When I am alone, I like to
a._____watch TV b._____listen to the radio c._____take a walk

28. When I am alone, I like to
a._____look at magazines b._____listen to music c._____play games

29. I enjoy
a._____painting b._____music c._____sports

30. In my free time, I like to
a._____draw something b._____talk to somebody c._____make something

31. In class, I sometimes
a._____look at a magazine b._____listen to my friends c._____play with my pencil

32. In class, I sometimes
a._____look out the window b._____talk to someone c._____move around in my chair

33. I learn best
a._____in class reading/writing b._____in class discussions c._____in class projects

34. I like to learn English by
a._____watching a video b._____listening to a tape c._____doing a role play

35. In class at school, I like
a._____colorful textbooks b._____interesting lectures c._____active lessons

36. I like
a._____movies b._____music c._____making things

TOTAL

a.______________ b.______________ c.______________

Appendix 3

What is your learning style? version three

Name ______________________________ Student Number_____________________

There are 3 answers in each line. Write number "3" next to the answer you like best. Write number "2" next to the answer you like second best and write number "1" next to the answer you like third best.

1. When I am lost, I like to

a._____look at a map b._____ask someone directions c._____go with someone

2. I like to

a._____see the words b._____say the words c._____move my hands with the words

3. I learn best when I

a._____see something b._____hear something c._____touch something

4. I like

a._____watching animals b._____listening to animals c._____touching animals

5. I learn best when I

a._____look at something b._____say something c._____touch something

6. To find a new place, I say

a._____"draw a map for me" b._____"tell me the way" c._____"take me there"

7. At the zoo, I like

a._____looking at the animals b._____hearing the animal sounds c._____petting the animals

8. I learn best by

a._____reading words b._____hearing words c._____acting with words

9. I like to

a._____write to friends b._____telephone friends c._____travel to friends

10. I like to contact friends by

a._____writing a letter b._____calling on the telephone c._____going to their house

11. I learn best by

a._____reading stories b._____hearing stories c._____acting stories

12. I like to

a._____look at the board b._____listen to the teacher practice c._____stand up and

TOTAL

a.______________ b.______________ c.______________

Appendix 4

Pair Correlation for LSQ v2

	Visual		Auditory		Kinesthetic	
pairs	r	p-value	r	p-value	r	p-value
16/21	.304	.0859	.357	.0406	.599	.0002
6/11	.555	.0006	.370	.0336	.696	<.0001
12/3	.110	.5437	.430	.0117	.382	.0277
*27/28	-.237	.1861	.258	.1484	-.251	.1605
*30/7	.171	.3436	-.707	.9691	.374	.0314
*26/9	.335	.0561	.143	.4299	.257	.1506
17/18	.827	<.0001	.387	.0253	.436	.0104
*2/29	.252	.1591	.363	0371	.082	.6512
*24/15	.268	.1329	.071	.6971	.247	.1675
*20/25	.039	.8289	-.298	.0927	-.309	.0802
*10/35	-.281	.1132	-.050	.7836	-.014	.9380
4/23	.534	.0011	.452	.0076	.520	.0016
8/13	.762	<.0001	.774	<.0001	.554	.0006
*22/19	.260	.1448	.406	.0183	.304	.0856
*31/32	-.138	.4478	.570	.0004	.041	.8202
*34/1	.259	.1460	-.094	.6052	.659	<.0001
*36/5	.023	.9009	.306	.0835	.579	.0003
*14/33	.221	.2177	-.344	.0493	.144	.4277

(Notes:) r = correlation * = pairs eliminated from version 2

Does It "Work"? Evaluating Language Learning Tasks

Rod Ellis
Temple University

Introduction

A quick look at the published work on materials evaluation (e.g., Cunninsgworth 1984; Breen and Candlin 1987; Skierso 1991; McDonough and Shaw 1993) reveals that it is almost entirely concerned with predictive evaluation. That is, it gives advice to teachers about how to conduct an evaluation of published materials in order to determine whether the materials are suitable for a given group of learners. This kind of evaluation is 'predictive' in the sense that it seeks to determine whether materials are likely to work in a specific teaching context. Valuable as this kind of evaluation is, it is not what I am concerned with here.

Instead, I want to consider how to carry out a retrospective evaluation of teaching materials. That is, I want to address how teachers can determine whether the materials they have actually used 'work.' It is my guess that although teachers frequently do ask themselves whether the materials they have selected or written 'work,' they generally answer this question impressionaistically in the light of their day–by–day experiences of using them. They rarely attempt a systematic and principled retrospective evaluation.

One obvious reason for this is the daunting nature of systematically evaluating the use of a complete set of materials (e.g., a textbook). This is an enormous undertaking, particularly if, as I shall shortly argue, the evaluation is to involve some kind of attempt to discover what it is the learners have learned as a result of using the materials. However, it may be easier to carry out retrospective evaluations at the micro-level by focussing on whether specific teaching tasks 'work.' My concern here, then, is with task evaluations.

What Does it Mean to Say a Task 'Works?'

A good starting point for a retrospective micro–evaluation is to ask what is means to say that a task has 'worked.' In fact, it can mean a number of rather different things. First, teachers might feel that a task has worked if they have evidence that the learners found it enjoyable and useful. The evidence might take the form of the teacher noticing that learners engage enthusiastically in performing the task or it might take the form of the students' responses to a post–task questionnaire designed to elicit how useful they felt it was. This kind of student–based evaluation is common and is the probably the basis for most teachers' judgements about the effectiveness of a task (see Murphy, 1993 for an example of a student–based task evaluation).

It is perfectly possible, however, that students enjoy doing a task and give it positive ratings in a questionnaire and yet fail to perform it successfully and/or learn nothing from it. It is also necessary, therefore, to consider two other types of retrospective evaluation; a response–based evaluation and a learning–based evaluation.

Richards, Platt and Weber (1985. p. 289) define a 'task' as 'an activity or action which is carried out as a result of processing or understanding language (i.e. as a response).' It follows that the effectiveness of a task might be determined by examining whether the 'response' of the learners is the same as the task was designed to bring about. This kind of evaluation constitutes a response–based evaluation.

A task may be more or less 'closed' or more or less 'open' according to the type of response asked for. In the case of tasks calling for verbal responses a fill–in–the–blanks grammar task can be considered 'closed' in the sense that there is only one set of right answers, while a free composition task can be considered 'open.' A non–verbal response may also be closed (e.g., a listening task that requires learners to fill in missing names on a map) or open (e.g. a listening task that asks learners to read a story and draw a picture of what they think the main character looks like). Now, it is obviously much easier to determine whether the 'response' learners make matches the one they were intended to make when the task is a closed one. Thus, teachers might feel the closed grammar and listening tasks outlined above have 'worked' if they observe that the students have filled in most of the blanks correctly and have been able to write down the missing names on the map. It is much more difficult to decide whether an open task has 'worked' as this requires teachers to identify criteria to evaluate whether the learners' responses are appropriate or not. For example, the students' response to the free writing task would need to be evaluated in terms of a set of criteria for effective writing (e.g., some kind of analytical marking scheme). The picture–drawing task would need to be evaluated in terms of the extent to which the students' pictures took account of the textual clues regarding the nature of the main character.

Thus, whereas the criteria for the evaluation of a 'closed' task are embedded within the task itself, the criteria required for evaluating an 'open' task are not. They are external to the task and, because they are usually not specified by the person who devised the task, they place a considerable burden on teachers' shoulders. This burden is notable because, in accordance with the dictums of communicative language teaching, many teachers are making greater use of 'open' tasks. It is my guess that many 'open' tasks are evaluated impressionistically. That is, teachers do not generally make explicit the criteria they are using to determine whether the learners' responses are effective or not.

Evaluating the effectiveness of a task in terms of whether the learners' responses are correct or appropriate constitutes what I call an internal evaluation. The evaluation is 'internal' in the sense that no attempt is made to ask whether the nature of the response required by the learner is a valid one: the evaluator simply assumes that the response required is valid and tries to establish whether the learners' actual response matches the response intended by the task.

Such an evaluation is, of course, limited because it is possible for a response to be correct or appropriate but still not be valid. It might be argued, for example, that a grammar task that requires learners to fill in the blanks with correct grammatical forms does nothing to promote the acquisition of these forms (see Krashen, 1982). It might also be argued that having students write free compositions does little to improve their writing skills. Furthermore, it is perfectly possible that a task fails to produce the intended response in learners and yet contributes to their development in some way (e.g., learners may fail to answer a set of comprehension questions on a reading passage correctly and yet learn a number of new words as a result of completing the task). In short, a task may be effective but invalid or it may be ineffective and yet valid.

A full evaluation of a task, therefore, calls for an external evaluation. It is possible to carry out an external evaluation theoretically (i.e., by determining whether the assumptions that task designers make when they design specific tasks are justified in the light on some theory of language acquisition or skill development). In this case, the evaluation is predictive in nature. To evaluate a task retrospectively calls for investigating whether a task actually results in any new language being learned or in the development of some skill. In other words, it requires teachers to determine empirically whether the assumptions about learning that task designers make when they design tasks are valid. This calls for a learning–based evaluation. It is, of course, noteasy to demonstrate that a task – whether 'closed' or 'open' – has contributed to language learning. One way might be to ask learners to note down what they have think they have learned as a result of completing a task (see Allwright, 1984 for discussion of 'uptake' as a measure of learning.)

To sum up, I have suggested that determining whether a task 'works' calls for different kinds of retrospective evaluations. A student–based evaluation provides information about how interesting and useful learners perceive a task to be. A response–based evaluation is internal in nature because it simply addresses the question 'Was the students' response the one intended by the designer of the task?' A learning–based evaluation is external in nature because it goes beyond the task itself by trying to determine whether the task actually contributed to the learners' second language proficiency.

The different kinds of evaluations – student–based, response–based and learner–based – call for different types of information and different instruments for collecting them. A full description of these information types and instruments is obviously needed but is not possible in this brief article.

Conclusion

The evaluation of language teaching materials has been primarily predictive in nature and has focussed on whole sets of materials. There is a need for more thought to be given to how teachers can evaluate the materials they use retrospectively on a day–by–day basis. I have suggested that this can be best carried out as a series of micro–evaluations based on the concept of 'task.' Such evaluations are likely to accord with teachers' own ideas of what evaluation entails.

Widdowson (1990) has argued the need for 'insider research,' by which he means that teachers should engage actively in trying out and evaluating pedagogic ideas in their own classrooms. Such 'action research,' he suggests, is essential to help teachers develop an increased awareness of the different factors that affect teaching and learning in classrooms. One way in which teachers can undertake 'insider research' is by conducting task evaluations.

Task evaluations, therefore, serve a double purpose. They help to determine whether particular tasks 'work' and, thereby, contribute to the refinement of the tasks for future use but, perhaps more importantly, they engage teachers as insider researchers and, thus, contribute to their on–going professional development.

References

Allwright, R. (1984b). Why don't learners learn what teachers teach? – The interaction hypothesis. In D. Singleton and D. Little (Eds.), *Language learning in formal and informal contexts.* **Dublin: IRAAL.**

Breen, M., & Candlin, C. (1987). Which materials? A consumer's and designer's guide. In L. Sheldon (ed.). Cunningsworth, A. (1984). *Evaluating and Selecting ELT Materials.* London: Heinemann.

Krashen, S. (1982). *Principles and practice in second language acquisition.* Oxford: Pergamon

McDonough, J., & Shaw, C. 1993. *Materials and Methods in ELT.* Oxford: Blackwell.

Murphy, D. (1985). Evaluation in language teaching: Assessment, accountability and awareness. In *Longman Dictionary of Applied Linguistics.* J. Alderson, J. Richards, J. Platt, & H. Weber, (Eds.) (1985). . London: Longman.

Skierso, A. (1991). Textbook selection and evaluation. In M. Celce–Murcia (Ed.), *Teaching English as a second or foreign language.* Boston, MA: Heinle and Heinle.

Widdowson, H. (1990). Pedagogic research and teacher education. In H. Widdowson. *Aspects of language teaching.* Oxford: Oxford University Press.

Communicative Oral Testing

Marion Delarche
Kanda Gaigo Daigaku

Nicholas Marshall
Kanda Gaigo Daigaku

Introduction

To teachers, testing often seems like something to be left to the "experts" who write thick books full of incomprehensible terminology. In our classrooms, and with our students, however, we often wish for better measures of their performance: ones that seem more in line with what we do in our classrooms than what is available on the professional testing market or what we tend to create for our classes.

What to Test

The first step in any description of an L2 testing device is a statement of what constitutes Language. For testing concerns, making this statement is essential because before one can make a test, one needs to have clearly in mind what is to be tested (Heaton, 1988).

Statements of what Language is have greatly evolved in the last few decades. Part of this evolution has been the change to a view of Language as the exchange of information, but we wish to take this a step further: we define Language as the exchange and further creation of meaning between interlocutors in a communicative way (Johnson, et al., 1995). What this means is that when interlocutors communicate, they not only exchange information, but together they build a set of information that, being unlike any other set of information between any other set of interlocutors, is a creation of new information. This is in complete discord with some other definitions of language, namely as a syntactic system that can be taken apart and "known" or as the correct answer from a set of four choices. Our definition instead recognizes that Language is made up of systems, both linguistic (morphosyntactic, phonological, etc.) and para-linguistic, and that the use of these systems is constrained by social, contextual, and numerous other factors.

Qualities Needed in a Test

Given the view of language described above, the qualities to look for in a test need to be defined. Wesche (1987) points out that a test needs to be, among other things, valid, pragmatic, focused on appropriateness and language in use, comprehensive of a variety of language functions, reliable, and feasible.

Combining Wesche's considerations with those above yields a long list of items to consider, so for reasons of space we will limit this discussion to the following: validity, reliability, schema-building, recognition of language components in scoring, and feedback.

Validity is often described in testing manuals as the single most important factor in testing, and indeed it is. There are many types of validity, but the one we are most concerned with for the purposes of this paper is construct validity. According to Heaton, "If a test has *construct validity* it is capable of measuring certain specific characteristics in accordance with a theory of language behavior and learning" (1988, p. 161). A test that is valid, then, can be said to assess what it claims to assess.

Discussions of validity are always accompanied by discussions of reliability, another of the most important factors in testing. Reliability has to do with the extent which a test is objective. If a test is completely reliable, then in theory, the same student taking the same test at the same time under the same conditions should score the same score. The reason validity and reliability are discussed together is because they seem to be inversely related: the more valid a test is, the less reliable it tends to be, and conversely, the more reliable a test is, the less valid it tends to be. Heaton (1988) points out, however, that in designing a test it is crucial to construct a test that is valid first and then to try to increase reliability—creating a reliable test and then trying to make it valid will not yield good results.

A third consideration that is not referred to nearly as often in testing literature is the importance of building schema before a test, both content and formal. Content schema is the background knowledge of a topic which the learner brings to a text with him or her, and has been discussed most in reference to L2 reading. The idea that in L2 teaching we need to help students build schema has been accepted for a decade, and it seems clear that the same should be true for testing: in not helping testees to build content schema, we risk testing them on what they know rather than on how successfully they manipulate language to exchange and create meaning.

Formal schema—the knowledge of the structure (in this case of a test) or of how to go about a task, can be just as important yet are often not considered. Again this poses a problem: if formal schema are not put in place beforehand, we risk testing not use of language, but testees' ability to figure out what is expected of them.

Another important factor in oral testing is the recognition of different components of language and the roles they play in communication. That is, we must recognize that the systems (linguistic and paralinguistic) of language can be teased apart to some degree for analysis; as well as recognizing that they are developed to different levels in different people. A test needs to distinguish where learners' strengths and weaknesses lie, especially since for teachers testing can be a teaching tool as well as an administrative tool.

Finally, the test needs to give testees useful feedback: it should describe a testee's level in each component, tell where strengths and weaknesses lie, and provide a basis for decisions about directions for future learning.

Norm-Referenced vs. Criterion-Referenced Tests

Most testing literature includes discussions of the differences between and uses of norm-referenced tests (NRTs) and criterion-referenced tests, so they will be discussed only briefly here (for more detailed discussion, see Brown, 1995 and others). A good example of NRTs and one that most of us are familiar with is the TOEFL. It can be administered easily to large groups, it is very reliable, and scoring involves a comparison between each testee and all of the others who have ever taken the test. This type of scoring, according to Wesche (1987), is less desirable when testing oral communication.

In our classrooms, we often use a different type of scoring for tests: criterion-referenced (CRTs). These kinds of tests are harder to administer to large groups and tend not to be as reliable; scoring involves setting a standard and comparing each testee to that standard independently of the other testees. If, for example, we teach our students a set of greetings, and then test them on their knowledge of those greetings, we compare each test to the standard of 100% learning of what was taught, regardless of how the other students have performed.

Professionally Marketed Tests

One NRT designed by Educational Testing Services as a test of oral skills is the SPEAK test. How well does it compare to the criteria set forth above? It is reliable, and although it is possibly a valid test of proficiency, is not a valid as a test of *communicative* proficiency—the only interaction involved is between the testee, a test booklet and cassette—communication strategies and knowledge of social constraints are not involved. Formal schema are not a problem for those who have taken practice tests or the real test at least once, but content schema are not built up at all from item to item. In marking the SPEAK test, scorers do refer to several components (Clankie, 1995) but the score given to testees does not reflect this breakdown.

There are also several oral communication tests available that involve criterion-referencing rather than norm-referencing. These include the American Council on the Teaching of Foreign Languages (ACTFL) oral interview and a variety of oral tests administered by University of Cambridge Local Examination Syndicate

(UCLES).

The ACTFL interview can also be assessed in terms of the criteria for testing described above. First, it is more valid as a test of oral communication than a test like the SPEAK test, because it involves interaction between interlocutors. However, one of the interlocutors is the tester, and the format is mostly question and answer. This, we feel, does not reflect a true-to-life pattern of interaction. As for reliability, the ACTFL interview enjoys quite a good rate, due mostly to painstaking care in training the testers. Schema present a problem, though: those who have not taken the test before do not have formal schema in place, and the probe part of the interview involves greatly varied (and sometimes bizarre) content. Scoring of the ACTFL interview involves placing students in one of several level bands (Omaggio, 1986; Nagata, 1995). Unfortunately, although the bands describe levels of ability in various sub-skills, the bands are not broken down into components, so that testers must give the same score for all components. As a result, feedback does not provide a description of testee strengths and weaknesses.

UCLES has developed a battery of tests for assessing oral skills, one example of which is the Cambridge Assessment of Spoken English (CASE). CASE consists of negotiation of a problem by testees, done first in pairs and then groups, or vice-versa. Scoring is done by means of a set of descriptor bands that have been broken down into several categories. Scores are assigned in each category and are then added for an overall score. As a test of communication CASE is highly valid, given the statement of Language above. UCLES as a professional testing organization, does its utmost to ensure high reliability rates. Again, for those who have taken the test at least once, formal schema are most likely in place, but no effort is made to build content schema before the task begins. As mentioned, scoring does include the use of a range of sub-skills or components, and so feedback does as well. As such, CASE fits fairly well our profile of a desirable test. Unfortunately, it is not available for classroom use.

Our Test Model

The test model proposed here fulfills, we hope, all the criteria for testing described above. A description of the test procedure follows. Each test, as described here, takes approximately twenty minutes.

In groups of three, students are given information sheets (see Appendix A—sample test materials) and presented with a problem to solve or a decision to make based on that information. First, however, they must complete an information gap task created by the existence of several blanks on each of their sheets. There are two kinds of gaps: those for which both of the other students can provide the missing information, and those for which only one of the others can provide the information. In this way, each testee is required to help in the exchange at least once, and then is provided with an opportunity to show willingness to provide information when not required to do so. Once all of the students have all of the information, the negotiation/ decision-making part of the test starts. Students are told that they must come to a joint agreement, and discussion begins.

The procedure described above is, we feel, a valid test format given the definition of Language set out in the "What to Test" section above. In exchanging information and negotiating a decision based on that information, testees must construct meaning among themselves in order to complete the task. The reliability of this test, however, remains uncertain. It has not been piloted or subjected to statistical analysis; again, we note Heaton's (1988) statement that validity needs to be of higher concern than reliability. We also expect that scoring with the aid of well-defined descriptor bands such as we will propose directly increases reliability.

As mentioned, schema-building also needs to be of concern in testing, and we find that the model proposed here accomplishes this. Formal schema are addressed by means of a practice version of the test, done as an ungraded class exercise, with students working in the same group in which they will be taking the scored test. The practice test follows the same format and procedure, but uses a different set of information and requires a different decision be made. Content schema, including key vocabulary, are built through an unscored pre-test exercise (see Appendix A) in which each student prioritizes a list of factors to consider in the decision.

The scoring bands used in this test (see Appendix B—oral test descriptor bands) are based on those used in the Kanda English Proficiency Test (KEPT), now in use at Kanda University of International Studies, and include four separate categories. The number of categories used is to some degree arbitrary, and is one of the problems that arises in creating a scoring system that accounts for the conflicting interests of rigor and ease of use. While three of the four categories in this scheme focus on linguistic factors, the fourth describes interactive aspects that fall outside systems of lexis, grammar and phonology. This is very important if we are to recognize Language as Halliday (1985) puts it, as a "...meaning potential system which is negotiated in interaction with others." We need to include these non-formal skills in scoring because by using communication strategies effectively those with poor formal language skills may interact effectively with others.

The scoring scheme used in this test is also used to provide feedback to students (see Appendix C—student oral score report). Note that the descriptors have been re-cast to avoid applied-linguistics jargon. Again, receiving scores in several categories with clearly described behaviors, students can see where their strengths and weaknesses lie; not only linguistically but also in terms of interactional skill.

Conclusion

There are many more issues involved in testing than can be discussed in these few short pages, and this is especially true in the case of oral testing. The testing format suggested above is by no means a final solution to the problem of how to accurately test communicative ability; but it may serve as a useful addition to the battery of tests we, as teachers, have at our disposal.

References

Brown, J. D. (1995). Differences between norm-referenced and criterion-referenced tests. In J. D. Brown (Ed.), *Language testing in Japan* (pp. 12-19). Tokyo: Japan Association for Language Teaching.

Clankie, S. (1995). The SPEAK test of oral proficiency: A case study of incoming freshmen. In J. D. Brown (Ed.), *Language testing in Japan* (pp. 119-125). Tokyo: Japan Association for Language Teaching.

Halliday, M., (1985). *An Introduction to functional grammar.* New York, NY: Edward Arnold.

Heaton, J. B. (1988). *Writing English language tests.* New York, NY: Longman.

Johnson, F., Delarche, M., Marshall, N., Wurr, A. & Edwards, J. (in press). Changing teacher roles in the foreign-language classroom. *Proceedings of the 1995 Chulalongkorn University conference on education.* Bangkok: Chulalongkorn University.

Nagata, H. (1995). Testing oral ability: ILP and ACTFL oral proficiency interviews. In J. D. Brown (Ed.), *Language testing in Japan* (pp. 108-118). Tokyo: Japan Association for Language Teaching.

Omaggio, A. (1986). *Teaching language in context.* Boston: Heinle & Heinle.

Wesche, M. (1987). Communicative testing in a second language. In M. Long & J. Richards (Eds.), *Methodology in TESOL* (pp. 373-394). New York, NY: Newbury House.

APPENDIX A—SAMPLE TEST MATERIALS

PRE-TEST EXERCISE

If you were choosing a place to study English in another country, what would be the most important things to consider in your decision? Look at the list below, and number the items in order of importance. Use "1" for the most important, and "7" for the least important.

_____Where the school is located
_____How much the school costs
_____How much it costs to live in that city
_____How much air fare is to that city
_____How many Japanese students the school has
_____What the weather is like in that city
_____Something else (what? ______________
____
________________________________)

TEACHER COPY

You are going to study English in North America for one year. Choose the school you will go to. You must all agree to go to the same school.

	UNIVERSITY OF PITTSBURGH [USA]	UCLA (LOS ANGELES) [USA]	UNIVERSITY OF VANCOUVER [Canada]
COST	¥700,000 per year, plus books	¥680,000 per year, plus books	¥650,000 per year, plus books
PERCENTAGE OF STUDENTS THAT ARE JAPANESE	9%	64%	32%
WEATHER	comfortable in Spring & Fall, hot in Summer, cold in Winter	mild in Winter, hot in Summer, comfortable in Spring & Fall	mild, but rainy in all four seasons
COST OF LIVING	not bad--a one-bedroom apartment is about ¥40,000 per month	high--a one-bedroom apartment is about ¥60,000 per month	not bad--a one bedroom apartment is about ¥50,000 per month
ROUND-TRIP AIR FARE	¥90,000	¥58,000	¥60,000

STUDENT 1

You are going to study English in North America for one year. Choose the school you will go to. You must all agree to go to the same school.

	UNIVERSITY OF PITTSBURGH [USA]	UCLA (LOS ANGELES) [USA]	UNIVERSITY OF VANCOUVER [Canada]
COST	¥700,000 per year, plus books	¥680,000 per year, plus books	¥650,000 per year, plus books
PERCENTAGE OF STUDENTS THAT ARE JAPANESE	9%	______	32%
WEATHER	comfortable in spring & fall, _________ , cold in winter	mild in winter, hot in summer, comfortable in spring & fall	mild, but rainy in all four seasons
COST OF LIVING	not bad--a one-bedroom apartment is about ¥40,000 per month	high--a one-bedroom apartment is about ¥60,000 per month	not bad--a one bedroom apartment is about ¥______ per month
ROUND-TRIP AIR FARE	¥90,000	¥58,000	¥60,000

STUDENT 2

You are going to study English in North America for one year. Choose the school you will go to. You must all agree to go to the same school.

	UNIVERSITY OF PITTSBURGH [USA]	UCLA (LOS ANGELES) [USA]	UNIVERSITY OF VANCOUVER [Canada]
COST	¥700,000 per year, plus books	¥680,000 per year, plus books	¥_______ per year, plus books
PERCENTAGE OF STUDENTS THAT ARE JAPANESE	9%	64%	32%
WEATHER	comfortable in spring & fall, hot in summer, cold in winter	mild in winter, hot in summer, comfortable in spring and fall.	mild, but rainy in all four seasons
COST OF LIVING	not bad--a one-bedroom apartment is about ¥40,000 per month	high--a one-bedroom apartment is about ¥60,000 per month	not bad--a one bedroom apartment is about ¥_______ per month
ROUND-TRIP AIR FARE	¥_________	¥58,000	¥60,000

STUDENT 3

You are going to study English in North America for one year. Choose the school you will go to. You must all agree to go to the same school.

	UNIVERSITY OF PITTSBURGH [USA]	UCLA (___________) [USA]	UNIVERSITY OF VANCOUVER [Canada]
COST	¥700,000 per year, plus books	¥680,000 per year, plus books	¥_______ per year, plus books
PERCENTAGE OF STUDENTS THAT ARE JAPANESE	9%	_____ %	32%
WEATHER	comfortable in spring & fall, hot in summer, cold in winter	mild in winter, hot in summer, comfortable in spring & fall	mild, but rainy in all four seasons
COST OF LIVING	not bad--a one-bedroom apartment is about ¥40,000 per month	high--a one-bedroom apartment is about ¥60,000 per month	not bad--a one bedroom apartment is about ¥50,000 per month
ROUND-TRIP AIR FARE	¥90,000	¥58,000	¥60,000

APPENDIX B—ORAL TEST DESCRIPTOR BANDS

	Pronunciation/ Fluency	Grammar	Vocabulary	Communicative/ Interactive Skills and Strategies
5	Rarely mispronounces. Accurate use of l/r, b/v, th, f. Speech flow rarely interrupted by difficulty in selection.	Uses high level discourse structure. Occasional errors.	Wide range from text with appropriate use plus appropriate lexis from outside text.	Confident and natural, responsive to others, aware of turn-taking. Asks others to expand on views. Body language natural.
4	Overall accurate pronunciation L2 influence does not impede comprehension for NS. Occasional interruptions in fluency.	Full range of basic structures; mistakes do not interfere with communication.	Lexis from text sufficient for task. Usually appropriate usage. Some lexis from outside text.	Responds appropriately to others. Needs prompting only occasionally. Can change topic.
3	Pronunciation often faulty but intelligible. Speech flow hesitant, some paraphrasing.	Meaning expressed in accurate simple sentences. Complex grammar avoided.	Lexis from text usually adequate for task. A little lexis from outside text.	Responds to others, usually does not require prompting. Can communicate main ideas. Sometimes uses repair and clarification strategies.
2	Frequent errors in pronunciation, sometimes unintelligible utterances Overall comprehensible. Speech broken except for routine expressions.	Errors frequent but intelligible to NS accustomed to NNS patterns.	Lexis from text adequate for simple communication only.	Requires continued prompting, otherwise silent. Does not initiate interaction. Difficulty following shifts of topic.
1	Pronunciation largely unintelligible. Both inaccurate and inconsistent. Very long pauses in selection of items.	Fragmented phrases. Single words.	Little lexis retained from text. Inadequate for simple communication.	No interaction without direct prompting. Speech very hesitant, not associated with what others say.

APPENDIX C—STUDENT ORAL SCORE REPORT

	Pronunciation/ Fluency	Grammar	Vocabulary	Communicative/ Interactive Skills and Strategies.
5	You made almost no mistakes, even with l/r, b/v th or f. You almost never had to stop to think about what to say.	You made a few mistakes, but you were using varied and complicated grammar and phrasing.	You used a lot of the vocabulary from the information page, and also used a lot of other vocabulary that you know.	You spoke confidently and naturally. You responded to other people or asked them to say more, and showed you understood the social rules of conversation. Your body language was natural.
4	You have an accent, but a native speaker would have no problem understanding you. Sometimes you had to stop and think a little about what to say.	You used a lot of different grammatical structures, and your mistakes didn't cause communication problems.	You used the words from the information page correctly and also used some other words you know correctly.	You almost always talked even if no one asked you to, and you were able to follow changes in the conversation.
3	You made mistakes, but it wasn't a problem understanding you. You spoke a little slowly, and sometimes you had to try to say something again in a different way.	Your grammar was correct but you only used simple grammar. You didn't use complicated grammar.	You used the words from the information page, and also used some other words that you know.	Most of the time, you talked even if no one asked you to. You communicated the main ideas although sometimes you had to explain more carefully to make your ideas clear.
2	Sometimes your pronunciation was difficult to understand, but generally it was okay. You spoke very slowly (except for common phrases).	You made a lot of mistakes, but someone who is used to non-native speakers could understand you.	You used the words from the information page, but you didn't have enough vocabulary to express complicated ideas.	You often didn't speak until someone asked you to, and you didn't start the conversation. You didn't seem to always follow changes in the conversation.
1	Your pronunciation was very hard to understand. You paused too long to think about what to say.	You didn't use sentences—only phrases or words.	You didn't use the words from the information page. Your vocabulary seemed too small for communication.	You didn't try to speak unless someone asked you to. Often what you said wasn't related to what other people said.
your score				

The Evaluation of Gestures in Non-Verbal Communication

Barry O'Sullivan
Okayama University

Introduction

If it is accepted that non-verbal (NV) communication strategies are an important element of our social interactions (Birdwhistle 1970, 1974; Morsbach 1973; Rubin 1982; Pennycook 1985; Seward 1968), it follows that we should accept the need to either explicitly teach them, or attempt to raise the learners' consciousness of them (Al-shabbi 1993; Soudek and Soudek 1985). This done, we should also accept the need to either test the learners' knowledge of such strategies or to in some other way quantify their ability to manipulate them in their communicative interactions (O'Sullivan, 1995b)

To date there have been no widely accepted efforts made to do just this, though there are recent studies (Jungheim 1995; O'Sullivan 1995a, 1995b) in which descriptions are given of attempts to generate tests which focus on the topic of the testing of non-verbal strategies. Though developed independently the tests share many characteristics, which are described below. However, before looking at these tests it is first necessary to make clear what we mean when we are discussing the area of NV communication. In order to do this, the models presented by Morain (1987) and Pennycook (1985) will be outlined.

Descriptions of the NV Channel of Communication

The most obvious difficulty one encounters when describing the NV channel of communication is its sheer complexity. The first example, Morain's (1987, p. 119) is aclassification of what she saw as a simplified outline of "the non-verbal aspects of communication":

1. Body language: comprising movement, gesture, posture, facial expression, gaze, touch, and distancing.
2. Object language: including the use of signs, designs, realia, artifacts, clothing, and personal adornment to communicate with others.
3. Environmental language: made up of those aspects of colour, lighting, architecture, space, direction, and natural surroundings which speak to man about his nature.

Though 'simplified' to the extent that each part is presented in broad definition, with no attempt to describe elements such as gestures in detail, even to the casual observer the above model is extremely wide-ranging. In terms of the language classroom the detail entered into here makes it of little practical use. Even when we look at the first level, that of 'body language,' it becomes patently obvious that it would be a practical impossibility to try to teach, let alone test, all of the elements in a normal language learning/testing situation.

In contrast to Morain's description, Pennycook (1985) focuses on the area of body-language only, and seems to provide us with a more practically useable format. However, while this appears to neatly categorise the area it does little to unravel its complexity:

1. Kinesics: Body movements, both deliberate and subconscious.
2. Proxémics: Private/Public domain, space judgements.
3. Paraverbal features: Non-lexical aspects of speech communication.

As with the Morain model, Pennycook focuses on broad definitions of the elements of the three principal facets, again making the model extremely wide-ranging and of little pedagogic use as it stands. Thus, in order to more fully understand non-verbal communication (NVC), and by implication make it more 'useful' to the language teacher and learner, we must attempt to more adequately describe it.

There are two very important points that can be made, having given these descriptions even a brief examination. These are:

1) The area of NVC is wide-ranging, complex and, to date, relatively unexplored, and
2) when we talk about gestures, or 'body movements,' we are, in fact, focusing on a very narrow aspect of NVC.

Tests of NVC Ability

Jungheim's two-pronged exploration of the subject saw him expand on Bachman's (1990) framework by adding what he describes as a "three-part nonverbal ability component" (Jungheim 1995, p. 150) comprising textual, sociolinguistic and strategic abilities. From this theoretical position he then proceeds to first outline his "Gestest" a 23-item test in which subjects were asked to respond to videotaped gestures — shown without sound — by correctly identifying their 'meaning' from a multiple-choice format. The test, which appears to have been methodically prepared, with numerous pilots and item description analysis used to come up with the final version employed in the study, generated impressive reliability statistics (he reports a Cronbach a coefficient of .75).

The second measure described by Jungheim was his attempt to use specially designed rating scales, which he called the NOVA Scales, to evaluate a learner's nonverbal strategic ability by enumerating their use of "head nods, gaze direction changes, and [hand] gestures" (Jungheim 1995, p. 157) in role play tasks which were videotaped and scored by trained raters.

O'Sullivan (1995a, 1995b) describes a study in which a similar test was created, though using just eight gestures, due to the effort to satisfy the cultural requirements of North American and British/Irish speakers of English. In attempting to look at the production and recognition of gestures this test consisted of two sections. In the first, the learners were asked to look at a gesture (embedded in a soundless videotaped scene performed by a team of native speakers of English (NSEs)— as opposed to Jungheim's (1995, p. 154) single North American female performing the gestures "while seated" — and then to identify its possible meaning from a multiple choice format, whose distracters were obtained from pilot test replies. In section two, the learners observed videos of interactions between NSEs — again without sound — these were cut just as a gesture was about to be made. The learners were then asked to perform a gesture which would 'fit' the cut-off point while transmitting a given meaning, this was given to them by means of a Japanese flash card. These performances were video taped and scored by trained raters. The small number of gestures, and the relatively small sample (n = 21) appear to have been among the factors which lead to the extremely low overall reliability scores observed (r = 0.423), though as can be seen below there were other difficulties.

Difficulties With the Tests

Both of the tests described above suffer from a number of serious drawbacks. For example, Jungheim's Gestest suffered from difficulties with the translation into Japanese of the intended meaning of gestures which had been originally written in English, as did O'Sullivan's test. The example mentioned by Jungheim (ibid., p. 157) was that of the expected response "I'm tired." intended to refer to "tired as in sleepy" being incorrectly translated as *tsukareta* , while the correct translation, *nemui*, was included as a distracter. Though the error was identified in time, the incident highlights the real difficulty of translating the intended meaning of a non-verbal cue from one culture into the written language of another.

In addition, O'Sullivan found that some of his raters accepted gestures that were seen as ambiguous by others, while Jungheim was forced to employ additional ratings when some gestures received widely differing scores — a likely indication that both tests suffered from this same malady. In terms of rater reliability there are two points to be made:

1) O'Sullivan's use of a vague concept of the 'acceptability' of gestures to the raters, by employing an holistic judgement appears to have been too subjective. This would certainly account for the poor inter-rater reliability obtained in his study.
2) Jungheim's NOVA scales, in offering just four levels of acceptability of an extremely limited number of very clearly observable items seems to offer a somehow naive or simplistic view of the situation. This is especially true when we consider the description of non-verbal channels of communication offered by Morain (1987) and that of body language from Pennycook (1985). The narrow bands described may also account for the high reliability scores he reported.

Some more fundamental problems become obvious when we consider the descriptions of NVC presented earlier. Though O'Sullivan was simply investigating the possibility of developing a test, Jungheim set out to develop a test which would act as a research tool to comprehensively examine the area. The small number of gestures either study identified are obviously not a representative sample of the elements of the descriptions offered by Morain and Pennycook, either in terms of the models as a whole or even of the single category of 'body movement' or 'gesture' and do not offer the examiner a sufficient basis on which a test could be drawn up even when all are included in every test — remembering that a smaller number of items on a test reduces its chances of generating acceptable reliability figures.

The method employed in both studies in presenting the gestures (using video without sound) cannot be seen as being authentic, when we consider that gestures require different degrees of required verbal and/or nonverbal input. By this it is meant that there are gestures which require; no spoken input, such as a victory sign, some degree of nonverbal input, such as the 'minimal responses' described by Zimmerman and West (1975, p. 108) an optional verbal input, such as a head shaking 'no,' or a combination of gestures/movements in a specific context to clarify the intent, for example a smile from a police officer when asking for your driving license does not necessarily mean that the officer is happy to see you.

Even where a gesture does not require verbal input; when it occurs it in some way changes the resulting message, for example either softening or intensifying it.

It is also clear that Jungheim's decision to use a seated gesticulator failed to take sufficient note of the interaction of different elements of NVC, remembering that the elements included in the descriptions outlined above are not likely to occur in isolation, but that there is a strong interaction between them. This seriously questions the validity of his method.

Observations and Discussion

That the literature has, to date, emphasized the culture specificity of the NV channel of communication is important to the EFL/ESL class in that it highlights two areas of concern to the language teacher and student. These are that we are on one hand failing to give our students the skills necessary to perform genuine communicative acts, while simultaneously ignoring an area of possible conflict in the language classroom.

Due to the focus of the typical language classroom there is a real possibility that the message transmitted through the verbal channel will be distorted because the accompanying nonverbal signals are misinterpreted or misunderstood, causing potential conflict both in the 'real' outside world and within the walls of the language classroom (see Al-shabbi, 1993).

However, even though the majority of the studies mentioned here are more than ten years old, and all, in one way or another either stressed the importance of NVC education or provided suggestions as to how it might be taught, the topic has rarely been included in an internationally published language text or teacher's manual. Difficulties, such as which elements of NVC to teach and of the fact that the culture specificity of gestures, makes including them in texts written for an international market all but impossible, contribute to this present situation. For similar reasons the creation of a widely acceptable test, certainly along the lines of those described above, appears to be fraught with apparently insurmountable difficulties.

Using the NV channel can be seen as a form of communication strategy. In the same way that repetition, pausing, and word coinage etc. allow the interlocutor to manipulate the communication system, non-verbal strategies allow us to transmit and interpret meaning. While some tests of spoken language (i.e., the UCLES batteries) contain instructions given to rater/interviewers which raise their awareness of the learner's inclusion of a number of communication strategies, this area has not been systematically explored for NV communicative ability. However, some awareness of the 'environmental language' is displayed in the instructions given to the instructors in relation to the physical organisation of the interview room (UCLES, 1988, p. 2-3).

Yet another reason for the neglect of this area may well be the success of Bachman's (1990) model of communicative language ability (CLA) in coming to dominate both language testing and research over the past few years. While it is extremely important for us to have a valid base on which to theoretically ground our research, and the model provides, in Bachman's (1990, p. 82) own words, "a guide, a pointer ... to chart directions for research and development in language testing," there is some difficulty in using it as a theoretical basis for evaluating a learner's communicative performance. This lies in the fact that in concentrating on the verbal side it does not concern itself with the evaluation/assessment of competence in the NV channel, an argument also employed by Jungheim (ibid., p. 149-151). In describing his framework as a guide Bachman calls for further expansion of the model through empirical research, a movement which Jungheim has certainly begun for NVC competence, though it is clear that there is much to be done.

Though the possibility of developing tests which will indirectly test such competence is certainly appealing, it is as inappropriate to separate the non-verbal channel from its natural context of communication as it is to separate the verbal channel. Therefore, in as much as previous tests can be argued to lack validity for ignoring one important aspect of communication, such indirect tests will lack validity for the same reason. In addition, it is also clear that the 'meaning' applied to any gesture will rely on the context in which that gesture is produced. It is important to realise, therefore, that to remove a gesture from its natural environment is to remove from it all meaning.

It is therefore apparent that language researchers/testers should continue to explore the whole area of non verbal communication. In addition to descriptions such as that offered by Pennycook we need to carefully study the individual elements of kinesics, proxemics, and paraverbal features so that we more fully understand their interactions, both among themselves and within the context any accompanying verbal or non-verbal communicative interaction.

At this point in time we simply do not know enough about the area to engage in test writing. It is therefore important to proceed with coordinated experimentation in order to create a validated working extension to the Bachman model. This achieved, it will be possible for researchers to examine the feasibility of including measurement of the NVC ability in existing tests of communicative competence. The conclusion that we should best proceed down this path is inevitable when we review the experience gained in failing to create a useable test of a learner's NVC ability when this ability is examined in isolation.

It is clear from the above discussion above that this writer has grave doubts about both his own efforts and those of Jungheim to write a reliable and valid test of a learner's NV competence. Additionally, there must remain a serious doubt whether such a test could or should be developed, even for research purposes, as the results generated tell us little or nothing of a learner's ability to accurately (or adequately?) interpret or produce signals on the non-verbal channel while engaged in a meaningful interaction.

References

Al-shabbi, A. E. (1993). Gestures in the Communicative Language Teaching Classroom. *TESOL Journal,* Spring, 16 - 19.

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Birdwhistle, R. L. (1970). *Kinesics and context: Essays on body motion communication.* Philadelphia, PA: University of Pennsylvania Press.

Birdwhistle, R. L. (1974). The Language of the Body: The Natural Environment of Words. In A. Silverstein (Ed.), *Human communication: Theoretical explorations.* Hillsdale NJ: L. Erlbaum Association.

Jungheim, N. O. (1995). Assessing the Unsaid: The Development of Tests of Nonverbal Ability. In J. D. Brown and S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 149 - 163). Tokyo: The Japanese Association of Language Teachers.

Morain, G. G. (1987). Kinesics and Cross-Cultural Understanding. In L. F. Luce and E. C. Smith (Eds.), *Towards internationalism* (2nd ed,) (pp. 117-142). Cambridge, MA: Newbury House.

Morsbach, H. (1973). Aspects of non-verbal communication in Japan. *Journal of Nervous and Mental Diseases,* 157(4), 262 - 277.

Neussel, F. (1985). Teaching Kinesics Through Literature. *The Canadian Modern Language Review, 41*(6), 1014 - 1019.

O'Sullivan, B. (1995a). The Production and Reception of Gestures in Non-Verbal Communication: A note on Evaluation. Paper presented at the RELC Conference, Singapore, April 1995.

O'Sullivan, B. (1995b). Evaluating Gesture Production and Recognition. *Bulletin of the Faculty of Education,* Okayama University, 99 (July), 231- 242.

Pennycook, A. (1985). Actions Speak Louder Than Words: Paralanguage, Communication, and Education. *TESOL Quarterly,* 19(2), 259 - 282.

Ruben, Brent D. (1987). Guidelines for Cross-Cultural Communication Effectiveness. In L. F. Luce and E. C. Smith (Eds.), *Towards internationalism* (2nd ed,) (pp. 117-142). Cambridge, MA: Newbury House.

Rubin, R. B. (1982). Assessing speaking and listening at the college level: The communication competency assessment instrument. *Communication Education, 31,* 19-32.

Soudek, M. & L. I. Soudek, L. I., (1985). Non-Verbal Channels in Language Learning. *ELT Journal, 39*(2), 109-114.

University of Cambridge Local Examinations Syndicate (1988). Instructions to Oral Examiners: First Certificate and Certificate of Proficiency in English. Internal Brochure. Cambridge: University of Cambridge Local Examinations Syndicate.

Zimmerman, D. H. and C. West (1975). Sex roles, interruptions and silences in conversation. In B. Thorne and N. Henly (Eds.), *Language and sex: Differences and dominance* (pp. 105-129) Rowley, MA: Newbury House.

Our Experiments in Oral Communication Tests

Shuichi Yonezawa

Nagano Prefectural Board of Education Consultant

OCA/OCB/OCC have been introduced into English lessons as part of the new English curriculum. One of the problems for English teachers is how to proceed with regular oral communication lessons with a textbook. Another is how to evaluate listening ability, speaking ability and oral communication ability. It seems that most of the teachers who are involved in these new subjects make a listening test in cooperation with AETs. Listening tests are likely to be accepted in some schools, partly because they have high administerability, and partly because they are a component of entrance examinations in some universities. Few teachers are trying to introduce speaking tests because they have problems of administration and objectivity. Our experiments in conducting speaking tests are just a small step toward the evaluation of oral communication in the present situation.

Subject

Forty first-year students of the English Course of Nakano-Nishi SHS participated in the

Speaking Test. They received one 50-min oral communication lesson per week which was team-taught with our AET from Canada, Kari McAlpine (She completed her teaching job in Japan in July, 1995. Now we have a new AET from Canada, David Kootnikoff). The textbook was Elementary LL English Course published by Taishukan. The usual lesson was made up of two parts. One part was for oral communication based on the textbook. The other part was for developing listening ability and global awareness by watching excerpts from BS news which I selected and recorded for the lesson. I have been trying to incorporate related reading and writing activities which are based on global education. The tests were conducted in June, November, and February, as a component of the three term-end examinations of the 1994 school-year.

Method

We studied various oral ability testings such as RAS (Royal Society of Arts) test, the ACTFL guidelines, the ILR (Interagency Language Roundtable) proficiency ratings, the ARELS (Association of Recognised English Language Schools) examinations, the BSM (Bilingual Syntax Measure), the Ilyan Oral Interview, the FSI oral rating system, the Clark four-scale system, the Jakobovitz-Gordon and Bartz rating system, and some other oral testings (Valette, 1977; Oller, 1979; Madsen, 1983; Byrne, 1986; Hughs, 1989; Heaton, 1990).

The First Oral Communication Test

The first oral communication test as the first-term examination was composed of two parts: 'Interpreter' and 'Talking' (see Appendix A).

The first part 'Interpreter' took one minute. Eighteen questions taken from the textbook were written on cards which were placed face-down on a table. Each student drew two questions, and handed one to the AET, one to the JTE. The student then acted as an interpreter between the AET and the JTE by translating the English question and its Japanese answer, and the Japanese question and its English answer. Grammar and accuracy were evaluated. The whole performance was recorded on video for later evaluation. In this part, accuracy was evaluated on the condition that one point was reduced for each mistake. The full mark of the first part was ten points.

The second part was 'Talking.' Students read a list of four topics and chose one. The four topics they could choose were: 1) Describe your morning routine, 2) Tell me about your family, 3) Tell me about your school life, and 4) Free choice. Students were to speak for two minutes, saying as much as possible. Some responses and questions from the AET were allowed during 'Talking'. According to our rating system, six things were evaluated: 1) the amount of information related (= the number of the sentences), 2) comprehensibility, 3) effort to communicate, 4) accent/pronunciation, 5) fluency, and 6) spontaneity. Students knew the process and evaluation scheme, but not the question, in advance. The whole performance was video recorded for later evaluation. The full mark of the second part was twenty-five points.

The AET and the JTE evaluated respectively, awarding thirty-five points maximum each for the whole test. These scores were totaled, for a maximum of seventy points then statistically converted to a ten-point scale in accordance Wit}l our school's evaluation system. We got 0.802 (p<.001) as inter-rater reliability.

The Second Oral Communication Test

The second oral communication test as the second term examination was composed of two parts: 'Appropriate Responses' and 'Free Speaking' (see Appendix B).

In the first part 'Appropriate Responses', students heard fourteen comments and responded to each comment in approximately fifteen seconds. Students were told in advance that they were supposed to give a natural answer and try to communicate meaning, and that they did not have to worry about grammar. The fourteen comments were picked out from their textbook. For example, students might hear a comment such as "Hi! I haven't seen you for a long time!" They might respond to it such as "Yes, it's been a long time." or "Hi, how are you?" The whole response was tape recorded in the LL room for later evaluation. A five-point scale was used to evaluate their performance, giving five points for each response. The full mark of the first part was seventy points.

In the second part 'Free Speaking', students were supposed to choose one topic and speak for one minute. They learned in the lesson about many global and environmental issues, based on the perspectives of global education. There were a lot of global issues we picked out: tobacco and second-hand smoke, The United Nations' peace keeping operation in Rwanda, nuclear energy, *waribashi*, world hunger, nuclear inspection in North Korea, trade friction between the USA and Japan, deforestation in Brazil and in other countries, and so forth. Students could talk about any other global issue other than the issues mentioned above if they were as an individual global citizen. According to our rating system, five things were evaluated: 1) length, 2) efforts to communicate, 3) amount of information, 4) level of English, and 5) understandability (= whether the judge can understand what issue each student is talking about). The whole response was tape recorded in the LL room for later evaluation. The full mark was thirty points.

The AET and the JTE evaluated respectively, awarding one hundred points maximum each for the whole test. These scores were totaled, for a maximum of two hundred points, then statistically converted to a ten-point scale in accordance with our school's evaluation system. We got 0.971 (p<.001) as inter-rater reliability.

The Third Oral Communication Test

The third oral communication test as the third term examination was composed of two parts: 'Speaking About the Cartoon' and 'Speaking About One Topic You Learned This Year' (see Appendix C).

In the first part, students were given one of four cartoons. They had two minutes to prepare. They had one minute to speak about the cartoon. They were supposed to refer to who, where, when, what, and how in their story, if it was necessary. 'What' was divided into four subcategories for us to put an emphasis on it: what he/she was doing, what he/she was thinking, what he/she was saying, and what he/she was feeling. According to each frame, their story was evaluated, in addition to the overall perspective of their story telling. The full mark of this part was twenty-eight points. The four cartoons we used for this part were originally for the pre-first grade test of the STEP. The whole performance was tape recorded for later evaluation.

In the second part, students were supposed to speak for one minute about what they learned this year. They could choose any topic that was discussed this year, and talk about it in detail, telling what they learned about that topic. They learned in the lesson about a lot of global issues, based on the perspectives of global education, as mentioned in the second oral communication test. In advance, students were given the grading frame of four things: 1) amount of information, 2) length of time talking, 3) whether it sounds like the student understands the topic, and 4) appropriateness of vocabulary. In addition, they were told that grammar was not graded so precisely, and that successful communication of ideas and their understanding of the particular global issue were important. We thought it did not matter if the issue the student picked out was the same as the one he/she chose in the second term oral communication test because it might lead to the better understanding of the issue and the more empowered communication of ideas. The full mark of this part was twenty-two. The whole performance was also tape recorded for later evaluation.

The AET and the JTE evaluated respectively, awarding one fifty points maximum each for the whole test. These scores were totaled, for a maximum of one hundred points, then statistically converted to a ten-point scale in accordance with our school's evaluation system. We got 0.879 (p<.001) as inter-rater reliability.

Results and Discussions

The First Oral Communication Test

Being time-consuming was one disadvantage. It took about four minutes for each student including change time, which amounted to one hundred and sixty minutes(= almost three hours). In addition, we needed almost the same amount of time for evaluation because we watched the whole performance on the video and counted the sentences for 'the amount of information related' item. Another disadvantage was that items for evaluation might overlap with each other. Rearrangement and integration was needed in selecting the evaluation items.

One of the advantages was that we could get the whole performance of the students by recording on the video. Another advantage was that by putting an emphasis on the amount of information related, we could approach one of the purposes of oral communication and have a highly objective evaluation in addition to the fact that JTE and AET evaluated respectively and got the total score, although we might not be free from some subjectivity.

Inter-rater reliability was 0.802 (p<.001), which was statistically significant. There was no significant difference between the two raters.

The Second Oral Communication Test

One of the disadvantages was that there was less naturalness in communication in the 'Appropriate Responses' because the student had no partner in the presence of him/her to talk with, so that the student had no chance to see and use any nonverbal communication such as facial expression, gesture, and eye contact. Another disadvantage was that 'Free Speaking' might be categorized into speech as one-way communication. It was not two-way communication, nor reciprocal. Thus, in this case, only one aspect of oral communication was evaluated. Reciprocity as the other aspect of oral communication was not evaluated.

The problem of being time—consuming was, to some degree, solved, because students were supposed to tape record their own performance according to the directions recorded in the tape in the limited time. It took about five minutes for each student. So the time needed to administer the second oral communication test was about five minutes. It amounted to about two hundred minutes in total for us to evaluate. But it was not so long or a laborious time. This was the first advantage in that the test had enough administerability. The second advantage was that we could have enough objectivity of evaluation as we used a five-point scale for the first part and five things to evaluate students' performance for the second part such as length, efforts to communicate, amount of information, level of English, understandability. In addition, the JTE and AET evaluated the same outputs respectively and got the total score. The third advantage was that appropriateness of verbal communication could be evaluated, though the time for the student to respond was limited and there was no non-verbal communication. The fourth advantage was that the student had an opportunity to speak about global issues, by expressing facts and their own opinions based on their learning and thinking in the lessons.

Inter-rater reliability was 0.971 (p<.001), which was statistically significant. There was no significant difference between the two raters.

The Third Oral Communication Test

One of the disadvantages was that both in the 'Speaking About the Cartoon' and in 'Free Speaking' the student had no partner to talk with in the presence of him/her, so that the student had no chance to see and use any non-verbal communication such as facial expression, gesture, and eye contact. Another disadvantage was that 'Speaking About the Cartoon' might be one-way communication. It was not two-way communication, nor reciprocal. Thus, in this case, only one aspect of oral communication was evaluated. Reciprocity or interaction as the other aspect of oral communication was not evaluated.

The problem of being time-consuming was, solved in this test, too, because students were supposed to tape record their own performance in the limited time. The time needed to administer the third oral communication test was about four minutes including the time for preparing how to construct a story. The student really spoke for two minutes out of four minutes in total. It amounted to about eighty minutes in total for us to evaluate. It was not so long or a laborious time. This was the first advantage in that the test had enough administerability. The second advantage was that we could have enough objectivity of evaluation as we gave points according to who, where, and what, for the first part, and we had four things to evaluate students' performance for the second part such as the amount of information, length, understandability, appropriateness of vocabulary. In addition, JTE and AET evaluated the same outputs respectively and got the total score. The third advantage was that appropriateness of verbal communication could be evaluated, though there was no nonverbal communication. The fourth advantage was that the student had an opportunity to speak about global issues, by expressing facts and their own opinions based on their learning and thinking in the lessons.

Inter-rater reliability was 0.879(p<.001), which was statistically significant. There was no significant difference between the two raters.

Conclusion

We have experienced three different types of oral communication tests. In the first test, we had a problem with the administration of the test, which we improved in the second test and the third test. But, instead of solving the problem of administration, we had the problem of unnaturalness of communication by tape recording their performances in that they had no real communication partner. Besides, in the speech type test and the story-telling type test, their performances had no reciprocity of communication as we had no device to insert our responses and questions to make them interactive and reciprocal. Thus, tape recording is a powerful way to solve the problem of administerability, but it can be a hindrance to reduce naturalness and reciprocity of communication.

We think that we cleared the problem of objectivity in evaluating students' performances from the first test in that we set some items necessary for analytic evaluation, and we had an appropriate inter-rater reliability. But we can safely say that we reduced naturalness and reciprocity as we tried to get objectivity by video recording and tape recording for the later analytic evaluation.

We may be able to improve these contradictory problems by adopting an interview type of oral communication test with an immediate evaluation whether it is holistic or analytic if we get used to evaluating students' performances. It is just an alternative way, so we would like to explore more alternatives for evaluating students' oral communication proficiency.

Reference

Byrne, D. (1986). *Teaching oral English.* Essex: Longman.

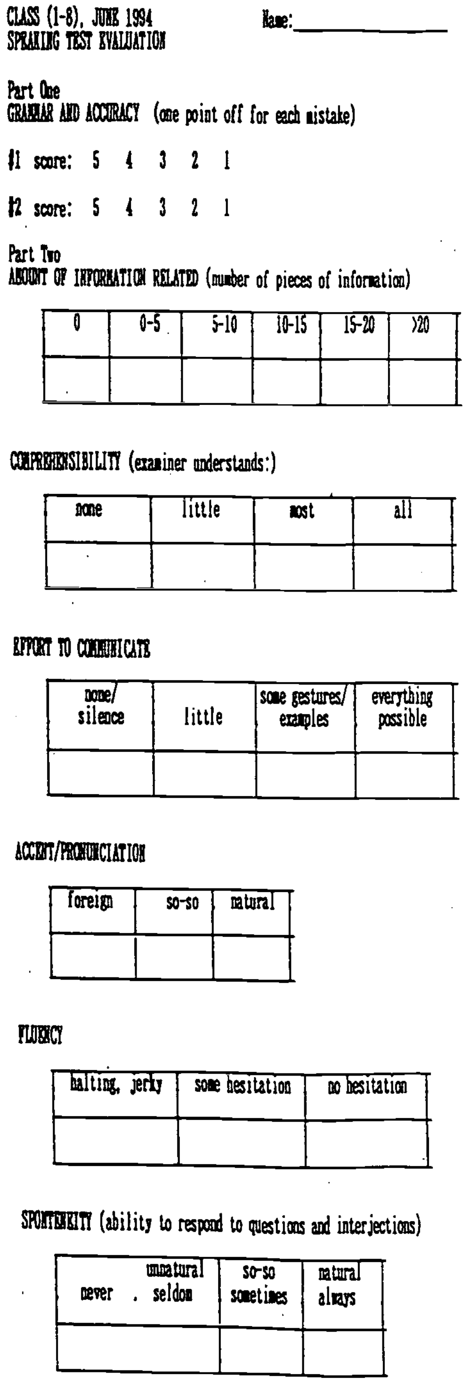
Heaton, J.B. (1990). *Classroom testing.* New York, NY: Longman.

Hughs, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Madsen, H. (1983). *Techniques in testing.* Oxford: Oxford University Press.

Oller, J.W.(1976). *Language tests at school.* London: Longman.

Valette, R.M. (1977). *Modern language testing.* New York, NY: Harcourt Brace Jovanovich, Publishers.



CLASS (1-8), JUNE 1994 Name:______________
SPEAKING TEST EVALUATION

Part One
GRAMMAR AND ACCURACY (one point off for each mistake)

#1 score: 5 4 3 2 1

#2 score: 5 4 3 2 1

Part Two
AMOUNT OF INFORMATION RELATED (number of pieces of information)

0	0-5	5-10	10-15	15-20	>20

COMPREHENSIBILITY (examiner understands:)

none	little	most	all

EFFORT TO COMMUNICATE

none/ silence	little	some gestures/ examples	everything possible

ACCENT/PRONUNCIATION

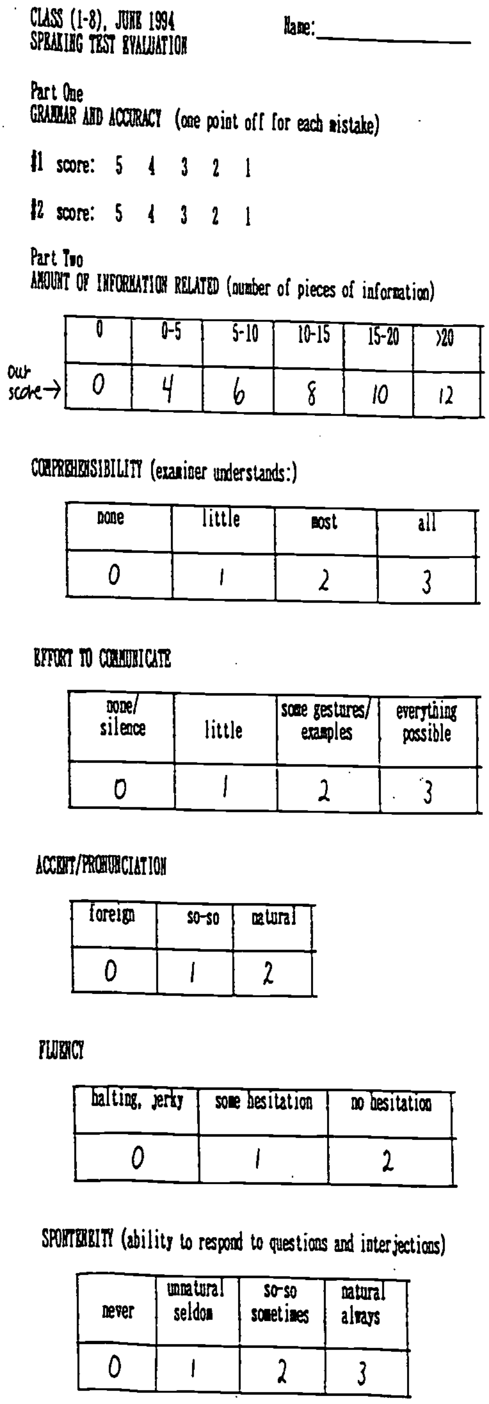
foreign	so-so	natural

FLUENCY

halting, jerky	some hesitation	no hesitation

SPONTANEITY (ability to respond to questions and interjections)

never	unnatural seldom	so-so sometimes	natural always



CLASS (1-8), JUNE 1994 Name:______________
SPEAKING TEST EVALUATION

Part One
GRAMMAR AND ACCURACY (one point off for each mistake)

#1 score: 5 4 3 2 1

#2 score: 5 4 3 2 1

Part Two
AMOUNT OF INFORMATION RELATED (number of pieces of information)

	0	0-5	5-10	10-15	15-20	>20
our score→	0	4	6	8	10	12

COMPREHENSIBILITY (examiner understands:)

none	little	most	all
0	1	2	3

EFFORT TO COMMUNICATE

none/ silence	little	some gestures/ examples	everything possible
0	1	2	3

ACCENT/PRONUNCIATION

foreign	so-so	natural
0	1	2

FLUENCY

halting, jerky	some hesitation	no hesitation
0	1	2

SPONTANEITY (ability to respond to questions and interjections)

never	unnatural seldom	so-so sometimes	natural always
0	1	2	3

CLASS (1-8), November 1994

SPEAKING TEST EVALUATION

Name:________________

Part One

Appropriate Responses

1	2	3	4	5	6	7	8	9	10	11	12	13	14

(5 point-scale)

Appropriate Responses Subtotal 70

Par Two

Free Speaking

how long?	< 30 sec.	< 45 sec.	< 60sec.
	2	4	6

effort	little	good	great
	2	4	6

quality

- amount of information	little	some	lots
	2	4	6

- level of English	easy	medium	advanced
	2	4	6

do I understand?	none	most	all
	2	4	6

Free Speaking Subtotal 30

TOTAL 100

PART 1

	Frame One	Two	Three	Four	
Who?					
Where?					
What?					
doing					
feeling					
thinking					
saying					
When?					
How?					
	6	6	6	6	24

overall:

do I know which cartoon was being described?	0	1
what happened before the cartoon?	0	1
what might happen next in the cartoon?	0	1
subjective mark, for exceptional work.	0	1

PART 1 TOTAL: ____
28

PART 2

amount of information
1 2 3 4 5 6 7 points

15 30 45 60 seconds
1 2 3 4 points

zero little ok good great student understanding
0 2 4 6 7 points

appropriate vocabulary words
1 2 3 4 points

PART 2 TOTAL: ____
22

TEST TOTAL: ____
50

Simulations: A Tool for Testing "Virtual Reality" in the Language Classroom

Randall S. Davis
Tokyo Foreign Language Business Academy

Introduction

Over the past two decades, a variety of non-traditional, humanistic teaching methods (e.g., Total Physical Response, the Silent Way, Suggestopedia, the Natural Approach, Community Language Learning, etc.) have been introduced to Japan in the hope that students will learn to speak English more fluently in their quest to the promise land of language mastery. Coupled with the ushering in of these methods, a new and greater emphasis on testing has emerged to the foreground.

Yet while many skills can be assessed using pencil-and-paper tests, oral proficiency "is widely regarded as the most challenging of all language exams to prepare, administer, and score" (Madsen, 1983, p. 147). Creating standard criteria of assessment, solving problems of administration, designing test items that resemble tasks in normal language use, and testing the complex and interlocking nature of language *and* skills in content-based courses are only a few of the logistic hurdles teachers must surmount in creating a sound testing instrument (Hughes, 1989; Littlejohn, 1990; McClean, 1995).

In Japan, the result has been that many teachers have resigned themselves to giving written tests instead; however, the concerns of creating a more enriched communicative environment for students and then assessing their language proficiency have led some to shift their attention to the use of simulations as a means of testing the language skill in action.

Simulations

The most common view of simulations is that they provide a way of creating a rich communicative environment (a representation of reality) where students actively become a part of some real-world system and function according to predetermined roles as members of that group. More important, however, is the notion that a simulation *becomes* reality and the "feeling of representivity fades" (Crookall & Oxford, 1990, p. 15), so much so that the world *outside* the simulation becomes, paradoxically, imaginary (see Black, 1995; Jones, 1982, 1985, 1987; Taylor & Walford, 1978, for a more detailed explanation of the mechanics of simulations).

The innate benefits of simulations include: (a) fulfill students' need for realism—a desire to "relate to life 'out there' beyond the classroom's box-like walls" (McArthur, 1983, p. 101); (b) increase student (and teacher) motivation, especially for those in EFL situations who might see English as a deferred need at best (Jones, 1982; Stern, 1980); (c) dismantle the normal teacher-student relationship so that students take control of their own destiny within the simulation, leading towards "declassrooming" the classroom (Sharrock & Watson, 1985); (d) help the learner confront and identify with the target culture (Oxford & Crookall, 1990); (e) reduce anxiety levels which is essential to language development (Dulay, Burt, & Krashen, 1982; Krashen, 1982); and (f) allow teachers to monitor the participants progress unobtrusively.

A Link Between Simulations and Language Assessment

As part of this movement, Littlejohn (1990, p. 125) suggests that "the use of simulations as a testing device is . . . an important development since it should be possible to replicate the situations in which learners will have to use the language." He also feels that this kind of replication "allows us to view not only the language product but also the process by which that language emerged" (ibid., p. 125). Whereas standardized methods give us insight on how the student might do in a real setting, "simulations will show us how the student actually performs" (ibid., p. 128; italics, the author's).

Let's Do Business: A Simulation Model for ESP Classes

Overview. To bridge this gap between simulations and testing, I have developed a task-based model at Tokyo Foreign Language Business Academy as part of an ongoing research project to evaluate the effects of simulation techniques in ESP classes, taking in account the need and desire to measure language proficiency (in this case, business English) at the intermediate level.

Design. Students are required to participate in a business simulation called "Let's Do Business" as part of the final evaluation near the end of the second year. This simulation deals with the rise of a travel agency called Fly Company from its inception through the research and development of a new sales promotion over a six-month period (which actually takes place during four consecutive class periods of 90 minutes each). I allocate each student the role of office manager, sales representative, or office clerk, and they are required to put into full use the language, behavioral, and business skills they have acquired during the past two years. In this case, I divide students into four branch offices of the company that are supposedly located in cities throughout Japan by partitioning the room into four sections, each equipped with a computer and printer, table and chairs, white board, phone, calculator, and access to a fax machine.

I make elaborate preparations to fulfill, what Jones (1982, pp. 4-5) terms, the three essential elements of simulations: (a) *Reality of function*: participants are assigned roles and are told they must fully accept them both mentally and behaviorally as if they were actually those people; (b) *Simulated environment*: a realistic setting constructed to enhance role-acceptance by utilizing a variety of realia, e.g., in this case, specially printed business cards, time cards, name tags, letterhead, technical support including computers and a fax machine, and memorandums; and (c) *Structure*: the whole action is built around a set of problems or tasks—not invented by the participants but rather evolve as the action progresses.

The groups are asked in a memorandum from the company president, William Johnson, to devise a new marketing strategy for domestic travel tours in Japan based on the results of a comprehensive survey of Japanese consumers' tastes and preferences. After analyzing the data, participants at each branch discuss their target market, decide how they are going to promote their services (e.g., television or radio spot, newspaper advertisement, direct mail, fliers, etc.), communicate their ideas and progress with the other branches by fax, phone, or mail, and then write and submit a proposal to the president.

In the end, our main goal is to provide some measure of both the *process* (how they approached the task orally in English by reviewing, organizing and weighing alternatives, deliberating over the information available to them, etc.) and the *product* (the proposal they draft demonstrating their English writing, computer, and reading skills).

Measuring the *Process*: Performance Checklists, Recordings, and Debriefing

The most challenging step is to evaluate the *process*. Three techniques that work well in tandem include a student-generated checklist, video or audio recordings, and a debriefing session.

1. *Job appraisal checklist.* One useful assessment tool I use is a student-created job appraisal checklist (see Appendix A, for one example) that, in reality, serves as a prop used by employees within this simulation as a way of measuring performance. Participants fill out this checklist based on whether they feel they fulfilled the duties as outlined in their job descriptions. The advantages of utilizing such a discovery approach are: (a) it empowers the participants with the know-how to evaluate their strengths and weaknesses without the constant feedback from an external evaluator; (b) its application is not limited to the classroom, but can be used later on the job; and (c) it satisfies the students' belief that their work should be fairly judged based on a system they clearly understand rather than be graded, in one of my student's opinion, "by a subjective scale created at the whims of the teacher."

Because I feel participant-reported responses often lack impartiality, I spend time training students how to be more objective by putting them in charge of writing the checklist as part of the regular coursework and then having them view past students on video engaged in similar business tasks and identifying positive models of the skills they want to acquire. Then, they practice evaluating each other in short role plays that resemble situations found in the simulation. At the same time, I take notes, record my own evaluations, and later discuss how my ratings coincide with those the students wrote down. My feedback at this point reinforces in their minds the validity and reliability of their own marks.

At the close of the simulation, the regional manager asks each participant to complete the job appraisal checklist before a year-end performance interview. The purpose of the interview, they are told, is to review their progress for possible promotion and pay raise in the near future. At this point, the simulation ends.

2. *Videotaping or tape recording*. Recording simulations can serve as a powerful tool for encouraging self-correction as well as student and/or teacher-initiated feedback. First, I try to position the camera so it will blend in with the surroundings without inhibiting students from assuming their roles in a more natural setting. I make sure the camera has become a regular fixture of the classroom weeks before (or months through repeated use) I carry out the simulation. By that time, students have accepted its presence and are not aware of whether it is rolling or not. Also, because four different meetings are going on simultaneously, I rotate the camera among groups to ensure that everyone appears on the video.

Furthermore, because tape recorders are always easier to come by and require less supervision, I set up a recorder in each office to tape the group's discussions. I connect the machine to a long extension cord and have the play button always on, so that by just plugging in the cord from outside their office, I can activate the recorder without participants conscious of when it is going or not.

3. *Debriefing*. The ultimate success of this simulation hinges on the efficacy of a wrap-up or debriefing session (together with the self-evaluation checklist and recordings) where students and the controller can openly discuss behaviors, outcomes, general language difficulties, and the contextual appropriateness of their language discourse. Because I, as the controller, do not take part in the simulation, I am able to look in as an observer without inhibiting students from assuming their roles.

Although there are several different approaches to debriefing (see Bullard, 1992), I hold a two-hour session the next class period, giving me time to reflect back on the simulation and organize my comments regarding students' behavioral or linguistic errors that were most apparent—and giving students a needed respite from such an intensive experience. Furthermore, as Bullard puts it, "the teacher has the chance to analyze the errors and to develop strategies for dealing with them at leisure rather than having to operate on the spur of the moment" (p. 64). Pedagogically speaking, this break has allowed me to view or listen to the tapes, record my observations, and prepare follow-up classroom lessons in the form of short role plays to reinforce areas that need improvement.

One simple technique for using the recordings in the debriefing is to write a checklist of listening or observation tasks. For example, I give students a checklist of the expressions studied in class for asking and expressing opinions in business settings, ask the students to watch the video, and check off the ones they hear, or see (in the case of certain non-verbal communication, e.g., gestures, facial expressions, paralanguage, etc.). Then, we come up with a group impression of how well students did.

Figure 1. Observation Task Sheet

1. Asking Opinions: What do you thin about...?
What's your opinion on...?
____

2. Expressing Opinions: If you ask me,...
In my opinion,...
____

3. Agreeing: You're eaxctly right.
Yeah.
That's how I feel!
I agree.
____

4. Disagreeing: I don't see it that way.
I don't agree.
I see what you're saying, but . . .
____

As the debriefing continues, I ask the participants to look at the remarks they made on the job appraisal form and critique their performance accordingly, checking to see if their own assessments concur with what they view on tape.

Measuring the *Product*: The Proposal

The second part of the evaluation deals with the *product*: the written proposal. I assign grades by looking at several specific criteria: (a) layout of the proposal (introduction, rationale, design, etc.), (b) mechanics (punctuation, spelling, and capitalization as studied in class), (c), content (organization, depth and breadth of arguments, and presentation of ideas), and (d) language usage (business terminology). I collect these proposals at the end of the simulation, and then score and return them. Each member of the group receives the same grade.

The Final Assessment: *Process* and *Product*

Ultimately, I meet with the participants individually to discuss comments and ratings on the checklist and to look over a copy of their proposal. We compare the results, and I give a final grade for the whole simulation project based on: (a) the student's own rating, 50%, (b) my assessment, 25%, and (c) the written proposal, 25%.

Study Design and Results

To determine both the effectiveness of the simulation and the value of the assessment tools used as viewed *by* the participants, I administered a short, written questionnaire comprised of four open-ended questions to 15 students in Japanese (to elicit more detailed comments), and these responses were then translated into English. [Those responses of particular interest have been cited here.]

The first question asked students to compare this simulation with other language activities in their other classes (e.g., dictation, skits, pair work, oral interviews, written tests, etc.). Eleven of the 15 students (S) regarded this technique more productive than other exercises they had experienced before:

S3: It [the simulation] was fun because the students were in control of the business rather than the teacher telling us what we should do next.
S5: It was a useful experience because the parts of the simulation didn't come straight out of a textbook.
S7: This activity combined what we practiced all year and what we will later need on the job.

The second question asked students whether they felt they had ample opportunities within the simulation to practice the skills studied in class:

S2: I like it because the phone conversations were not scripted by the teacher, but were created by the students out of a real need to communicate.
S11: Each thing we did was related to the next, so I had the chance to try many things at once.
S15: It simulated the pressures of the real thing and allowed me to see whether I had mastered my English or not.

The third question focused on whether the skills-assessment methods (checklist, videotaping, debriefing session, and proposals) were helpful in measuring students' abilities and provided enough diagnostic feedback to assist them in seeing their strengths and weaknesses for improvement.

S2: Talking to all the students together at the final meeting was good because I could see that other students had similar concerns and problems in English, and we could learn from each other.
S5: The evaluation sheet was useful because it helped me learn how to check my own ability.
S9: I enjoyed watching the video of the simulation because I could see myself using English. I always wondered if others could understand what I was saying.

The final question dealt with the overall design of the simulation and asked students how it could be improved. Of the 15 students, seven suggested no specific changes. The other eight students recommended modifications in format, timing, role allocation, and formal feedback. Some of these suggestions include:

S1: The first day was exciting, but as the simulation continued on over several classes, it lost some of its momentum.
S10: I wish more cultural issues in working with foreign companies would have been introduced.
S15: It would have been nice if there had been some foreign teachers acting as members of the staff to motivate and force us to communicate more in English.

Final Reflections

The results of the survey and my own observations have helped me chart a new course using simulations as the cornerstone of our program. One might question the plausibility of carrying out such elaborate simulations, considering the limitations of time and space, for example, while dealing simultaneously with weighty demands of classroom requirements already. Finding myself under the same constraints, I have slowly progressed from simple skits, to detailed role plays, to more involved productions over some time, giving myself time to digest and process this unique method of teaching and testing while gaining converts along the way. . . and the reward has encouraged me to push on.

Whatever the obstacles, the comments in the questionnaire have shown me that once students had tasted the benefits of simulation, their desires to learn improved considerably. Furthermore, the extent to which the students praised our efforts not only reflects how radically different this kind of approach still is in Japan, but how little simulations have permeated into the classroom although they have been the focus of discussion for many years in teacher-training circles. Finally, the students' responses seem to mirror the current state of affairs in many language-teaching settings: traditional methods of assessing oral proficiency do little to prepare the trainee for the realities and demands of life.

Since initiating the use of simulations as a pedagogical learning and testing tool in the classroom, my students and I have found a great sense of fulfillment and satisfaction in taking part in activities that innovative, pragmatic in nature, and fun. What Jones observed several years ago is just as, if not more, significant today: "The time seems to be ripe for extending their [simulations] use . . . particularly in the field of language assessment" (1982, p. 77).

References

Black, M. C. (1995). Entrepreneurial English: Teaching business English through simulation. *English Teaching Forum*, *33*(2), 2-9.

Bullard, N. (1992). Briefing and debriefing. In D. Crookall & R. L. Oxford (Eds.), *Simulation, gaming and language learning* (pp. 55-66). New York: Newbury House.

Crookall, D., & Oxford, R. L. (Eds.). (1990). *Simulation, gaming, and language learning*. New York: Newbury House.

Dulay, H., Burt, M., & Krashen, S. (1982). *Language two.* New York, NY: Oxford University Press.

Hughes. A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Jones, K. (1982). *Simulations in language teaching.* Cambridge: Cambridge University Press.

Jones, K. (1985). *Designing your own simulations.* London: Methane.

Jones, K. (1987). *Simulations: A handbook for teachers and trainers* (2nd ed.). New York: Nichols Publishing.

Krashen, S. D. (1982). *Principles and practice in second language acquisition.* Oxford: Pergamon.

Littlejohn, A. (1990). Testing: The use of simulation/ games as a language testing device. In D. Crookall & R. L. Oxford (Eds.), *Simulation, gaming and language learning* (pp. 125-133). New York: Newbury House.

Littlewood, W. T. (1981). *Communicative language teaching: An introduction.* Cambridge: Cambridge University Press.

Madsen, H. (1983). *Techniques in testing.* New York: Oxford University Press.

McArthur, T. (1983). *A foundation course for language teachers.* Cambridge: Cambridge University Press.

McClean, J. (1995). Negotiating a spoken-English-scheme with Japanese university students. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 136-148). Tokyo: The Japan Association for Language Teaching.

Onoda, S. (1995, September 18). Good testing methods a prerequisite for teaching. *The Daily Yomiuri*, p. 9.

Oxford, R. (Ed.). (1990). *Using and learning language through simulation /gaming.* Newbury Park, CA: Sage.

Sharrock, W. W., & Watson, D. R. (1985). Reality construction in L2 simulations. In D. Crookall (Ed.), *Simulation applications in L2 education and research.* Oxford: Pergamon.

Stern, S. L. (1980). Drama in second language learning from a psycholinguistic perspective. *Language Learning*, *30*, 77-97.

Taylor, J. L., & Walford, R. (1978). *Learning and the simulation game.* Milton Keynes: Open University Press.

Appendix A: Student-Generated Checklist

This assessment is based on the list of responsibilities and skills needed as a member of Fly Company. Use the following list to judge your own abilities and write other comments.

3 = Well done
2 = Fair - Needs improvement
1 = Unable to finish the work satisfactorily

1. I can use the computer to write letters/faxes/memos: 3 2 1
(format, addresses, punctuation, spelling, greetings and closings, envelope format, fax layout, abbreviations, speed, etc.)

____

2. I am able to answer the phone and take messages in English: 3 2 1
(answering the phone, asking for additional information, recording message correctly, responding quickly, etc.)

____

3. I work well with other employees in the office: 3 2 1
(helping others as a team and eager to do extra work when needed, etc.)

____

4. I am able to express my opinions clearly on important decisions: 3 2 1
(agreeing, disagreeing, persuading, asking questions, etc.)

____

5. I complete my assigned work on time: 3 2 1

____

6. I come to work on time: 3 2 1

Other:
____
____

Employee's Signature	Position	Date
Employer's Signature	Position	Date

Author Note

Correspondence concerning this article should be addressed to Randall S. Davis, Tokyo Gaigo Business Academy, 1-21-5 Morino, Machida-shi, Tokyo 194. The author can be reached at 0427-28-6751.

Evaluation of Listening-Focused Classes

Yoshinobu Niwa
Chubu University

Kazuo Iwata
Aichi Gakuin University

Introduction

This paper discusses the new curriculum of Aichi Gakuin University, the role listening-focused classes play, and presents a case study of a listening-focused class.

The New Curriculum of Aichi-Gakuin University and the Role of Listening Focused Classes

Why Were Listening-Focused Classes Introduced as a Core Subject?

The aim of the new curriculum, starting in 1994, was to respond to students' call for developing English proficiency in real situations. Two things accelerated its realization. One was the decision of the Ministry of Education (*Monbusho*) to move toward communicative English learning, and the other was to make summer language courses abroad successful. Aichi Gakuin students were not used to communicating with foreigners at all. They gave up easily more communication and were often content with the classroom English.

It suggested an important thing about this new curriculum. English teachers had to make students accustomed to communicating in English. How can they, especially students with lower levels of language attainment, manage it? For the new curriculum to respond to this question, it is needed, first of all, to provide all the students with listening and speaking classes. Generally speaking, Japanese university students have too little experience in listening comprehension and oral communication. According to the result of Questionnaire given between June and July, 1995, only 5-6% of the freshmen had classes of those kinds every week in the past (see Table 1).

Listening and oral communication were designed as one semester subjects, because students were only required to take three one-year classes although four classes were needed so that each skill-oriented subject could be taught intensively. One could have chosen to cut reading and writing instead, in consideration of what students lack. But most of English teachers thought that any more preference for listening/ speaking would be too radical. Moderate change was wanted.

Nevertheless, it was essential to give students a revolutionary image concerning the curriculum. Then it was decided to have all the students taught by native speakers of English who were to teach oral communication. Thus Japanese teachers of English were to teach listening-focused classes.

Why Have Listening-Focused Classes Been Taught by Japanese Teachers?

Two other reasons for separating listening from oral communication exist. One is futuristic: a design of collaboration between Japanese and native speaking English teachers in class activities. Any exercise of listening comprehension would be able to complement to oral communication and vice versa. Those two classes can be regarded as a sort of whole-year class.

The other reason is more serious. Even the moderate change in the new curriculum was really revolutionary to the Japanese teachers, because it increased the number of classes taught by native speakers of English: 44 in total for oral communication and English conversation, whereas only eight were necessary for English conversation before 1994.

Good reason for teaching practical English must be declared. Most Japanese teachers are probably at a great disadvantage unless they can tell students their own experiences in foreign countries about what makes it difficult to communicate and how they get over these difficulties. It should be personal, as there can be some truths hidden behind such experiences which English native speakers cannot notice because they are native. It is a sort of contribution to building up a method for teaching English to Japanese. And, generally speaking, Japanese teachers can contribute more in listening comprehension than in oral communication.

Are Students Content with Listening-Focused Classes?

The main aim of the questionnaire mentioned above is to know how students evaluate listening-focused classes. According to the results shown in Table 2 and Table 3, they are very successful. 42.3% of the students enjoy listening (Table 2) and 52.4% of them think listening-focused classes are useful as an initiation into communicative English lessons (Table 3).

The Difference in Students' Responses between Listening-Focused Classes and Oral Communication Classes: For Future Collaboration

The questionnaire has another aim: to investigate the difference in student responses between listening-focused and oral communication classes. Where does the difference, if any, come from? As the sum of the figures of 5 and 4 in Table 2 clearly show, oral-communication classes (62%) are more preferable than listening-focused classes (42.3%). It is well known that what students want most in university is native speakers' classes.

Such a preference by Japanese students seems closely related to the presumable crisis in the future for Japanese teachers mentioned above. But, according to the results of the questionnaire, the situation is not worse than expected. The number of the students who think listening-focused classes are useful (52.4%) is larger than that of those who enjoy them (42.3%). The negative answers also decrease from 13.1% in the question concerning students' enjoyment of the classes (Table 2) to 9.6% in the question concerning students' perceived benefits (Table 3). The result is also meaningful when compared to the response concerning oral communication classes, where the difference in the percentage of the negative answers between Table 2 (4%) and Table 3 (3.4%) is rather small. The number of positive answers even decreases from 62% in Table 2 to 58.8% in Table 3.

More interestingly, although the answer "so so" is most common (44.6%) to the question of how enjoyable listening is, the answer "useful" becomes the largest (43.2%) in Table 3 when the question comes to how beneficial it is. And the number of choices other than 4 ("useful") decreases, when compared to those in Table 2 (13.1% to 9.2% on 5; 44.6% to 38.0% on 3; 9.0% to 7.6% on 2; 4.1% to 2.0% on 1). It means the students who vary on how much they enjoy listening tend to agree more or less on its benefit.

On the other hand, the students who answer "so so" on the question how useful oral communication classes are (37.8%) is larger in number than those who answer the same on the question how enjoyable they are (34.0%). Correspondingly the answer "useless" in Table 3 (3.0%) is a little larger than the answer "not much" in Table 2 (2.6%).

Those results suggest that listening-focused classes can be roughly characterized by students as useful, and oral communication classes as enjoyable. Presumably students feel that native - speaking English teachers' classes, represented by oral communication classes here, are a kind of epicurean, fun-based English lessons, whereas Japanese English teachers' classes, represented by listening-focused ones, are a kind of stoic, continence-based English lessons. Some students even note in the questionnaires that they do not believe that language learning with much fun will be effective. The results of the questionnaire thus exemplify that the traditionally rigorous attitudes toward learning are still strong among young Japanese. Even the students who declare their liking for fun-based English lessons still seem to believe that language learning cannot be filled with fun.

Here are possibilities for Japanese English teachers' collaboration with native speaking English teachers. One can encourage students to study enjoyably or broad-mindedly, not to study rigorously. Or else one should bring home to students that listening classes are really useful although they are not fun. But all English teachers do not seem to recognize this enough. As many as 46.7% of the students in listening-focused classes cannot decide whether or not to take another listening class, according to Table 4. Standing apart from possibilities of other reasons, I would like to focus on this: they cannot decide because it would inevitably depend on the degree to which they are satisfied. It would probably also depend on what and how their teachers teach.

Table 1: Question: Did you have classes of the same kind in the past?

	Listening Classes	Oral Communica tion Classes
5: Every week	6.2	5.6
4: Sometimes per semester	10.7	11.8
3: Sometimes per year	18.5	15.9
2: Few classes in the past	36.4	39.5
1: No classes in the past.	28.2	27.2
	100(%)	100(%)

Table 2: Question: Did you enjoy this class?

5: Very much	13.1	24.6
4: Much	29.2	37.4
3: So so	44.6	34.0
2: Not much	9.0	2.6
1: Not at all	4.1	1.4
	100(%)	100(%)

Table 3: Do you think this class is useful?

5: Very useful	9.2	13.2
4: Useful	43.2	45.6
3: So so	38.0	37.8
2: Useless	7.6	3.0
1: Very useless	2.0	0.4
	100(%)	100(%)

Table 4: If the similar subjects are available, do you want to take them?

5: Definitely yes	8.0	18.3
4: Yes	30.7	39.8
3: Not decided yet	46.7	35.7
2: No	9.9	4.9
1: Definitely no	4.7	1.
	100(%) 1	00(%)

A Case Report: A Listening-Focused C lass

Niwa's Natural Method And Procedure

This method is a 'practice makes perfect' method. Listening to a story many times with the intention of understanding a story can get students to find the most appropriate method for themselves naturally. This method has nothing specific, such as paying attention to chunks, rhythm or pictures. The one important factor is to have the intention to understand a story and to try to predict a story. The rest of the psychological activities needed for listening is entrusted to individual linguistic instinct.

The procedure consists of listening and testing in each period. For listening, *Today' Japan, Listening-focused Exercises* by T. Yamazaki and Stella M. Yamazaki (1993) was used. Among 20 stories 6 were picked at random and 50 copies of two types of tests were prepared each time.

Listening should be natural and abundant. Each story is spoken at natural speed, probably with more than 160 wpm, and is rather a long story consisting of about 200 words. Each story is repeated 10 times in all.

Testing is of two types. The first is 3 true and false questions and 4 of multiple choice questions (this is referred to as Choice or C test). The second is a kind of dictation (or cloze), that is, 10 questions of filling in blanks with the appropriate words (this is referred to as Dictation or D test). In order to avoid students' preparing beforehand , the two types of tests were prepared each time and texts were not used at all. After collecting answer sheets for the Choice test, the answer sheets for the Dictation test were handed out. So the result of Choice test depends completely on listening experience. Further two teachers supervised during each test to prevent students from talking to each other about the answers.

Evaluation and Analysis

The following are the main points in the evaluation of this method and the analysis of the results of the two tests.

High Motivation

The first simple success of this class is that students devoted themselves to comprehension of the stories very seriously. Usually they talk to each other and are noisy in the class . The length of listening time is long and reaches as many as

50 minutes without a break. This simple exercise happens once in two weeks all through the term. Under such conditions they worked very hard, engaging in listening to the stories very seriously. This means that they had sufficient motivation to try to understand each story.

The Results of The D Test

The Dictation test did not show improvement. See scores and graphs in Figure 2. The coefficient of the Choice test and the Dictation test is very low. The highest coefficient is 0.47 between C4 and D4. The lowest is 0.14 between C1 and D1. This means that students did not improve in word-after-word, bottom-up listening processing.

The Result of The C Test (1)

In contrast with the Dictation test, students showed improvement in the Choice test each time (see scores and graphs in Figure 1). The number of students are 50 and maximum scores are 10 each time. The improvement is statistically significant between the beginning C1 and the last C6 (P= 5.714E-13). Improvement was even significant each time between C1 and C2 (P=0.0007019), between C2 and C3 (P=0.02), between C4 and C5 (P=0.02), but not significant between C3 and C4 (P=0.30) and between C5 and C6 (P=0.29). One can conclude that they made progress in top-down processing and predicting content.

The Result of The C Test (2)

In order to understand the reason for this improvement more, the results were analyzed, dividing the students to three groups: high level, intermediate level and low level (abbreviated HIL in the title of graphs below). The criterion for the level division depended on the scores of Choice test 1 + Dictation 1 (20 points). The average of high level group is 7.90, intermediate level group 5.6 and low level group 3.0.

Interestingly it was found that low and intermediate level students showed more improvement than high level students (compare the scores and graphs in Figure 4).

More clearly one can see the difference of improvement between these groups by comparing the results of the 1st test (beginning) and the 6th test (end) (see Figure 3). The low level group improved most from 2.21 to 5.78, then the intermediate level group from 3.84 to 5.64 and the high level group from 4.36 to 6.45. This means that improvement was made on the process of prediction or imagination rather than listening to each word, and as far as process of prediction goes, it seems that low level students have more room for improvement. It means even low level students can understand such an English story roughly and choose a correct answer , even though they do not understand each word, and probably the structure of each sentence. The process of prediction is a top-down process and is very important for everybody who engages in listening comprehension activities. Here Schlesinger' words in Rivers (1981, pp.161-162) strongly confirm this:

> In listening we may not bother to process most of syntax...we resort to the analysis of the syntax of the sound signal only when there is ambiguity or when, for some reason, we have not extracted a clear meaning from signal. If this is so, foreign-language learners need a wide recognition vocabulary for rapid comprehension, rather than a sophisticated knowledge of syntax.

However, this practice for listening has long been neglected in Japan, even in reading and writing. Teachers have emphasized translating Japanese or English sentences into English or Japanese sentences, accurately without grammatical errors. This traditional way of teaching has made students pay attention only to words or short sentences, neglecting the understanding of the meanings at a paragraph or story level. It seems that prediction is one of the important factors in understanding a story. Therefore, if the above assumed reason for this improvement is right , one can conclude that this listening focused class was successful and could supplement what has been neglected so far in Japan.

Students' Impression

The result of the C test (2) agrees with the response of each student to the questionnaire. Low and intermediate level students had an impression of more improvement than high level students (see Figure 5). Self evaluation is shown by scores: 1 (no progress), 2 (some progress), 3 (progress) and 4 (much progress). This result is partly confirmed by Iwata's questionnaire result.

Future Problem: Harmony between Top-Down and Bottom Up Listening

This listening class has produced a fruitful result. However the final goal of listening competence is far from being reached. This must include integration of top-down and bottom-up processes. Peterson (1961, p. 109) says, "This model of listening as an interactive process suggests a new integration of both global (top-down) and selective (bottom-up) listening in the class room." Much research so far has been done in Japan in order to improve the teaching of bottom-up processing in Japan. However, any concrete method to integrate both processes has not been suggested. Therefore all that was done this time is (1) to encourage the students to have the desire for understanding a meaning, (2) to have the competence of prediction about a story and (3) to have as much experience of listening to native speakers' speech as possible. It might be difficult to find any one method for harmonious integration of top-down and bottom-up processing . However, it is necessary and will be possible that an effective standard method for it will be found by repeating researches with patience.

References

Brown, G. (1977). *Listening to spoken English.* London: Longman.

Dunkel P., &Lim P.L. (1986). *Intermediate listening comprehension.*USA: Newbury House.

Kono, M., & Sawamura, F. (1985). *Listening and speaking—Atarashii kangaekata—Y*amaguchi Shoten.

Miller, G.A. (1956). The magic number seven, plus or minus two: Some limits on our capacity for processing information. P*sychological Review,* (3) pp. 81- 97.

Morley J., (1991). "Listening comprehension in second/foreign instruction." *Teaching English as a Second or Foreign Language.,* pp.81-106.

Peterson (1991). "A synthesis of methods for interactive listening. *Teaching English as a Second or Foreign Language.,* pp.106-122.

Rivers W., (1981). *Teaching foreign-language skills: 6. listening comprehension..* Chicago, IL: University of Chicago Press.

FIGURE 1

Multiple Choice Test Scores

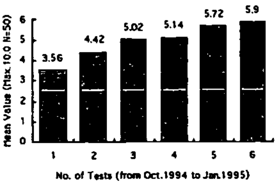


FIGURE 2

Dictation Test Scores

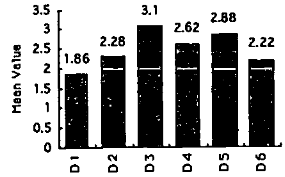


FIGURE 3

HIL 1st and 6th Choice Test Scores

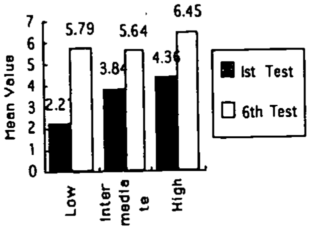


FIGURE 4

HIL C Test Result (from C1 to C6)

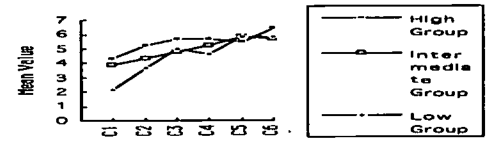
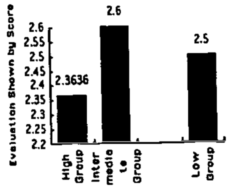


FIGURE 5

Self-Evaluation of Improvement



Interpreting Teacher and Course Evaluations

T. R. Honkomp
Chijushi Jogakuen University

Addressing the students' needs is an educational objective that most university instructors consider before the long-term planning of a semester course as well as before daily lesson planning and subsequent teacher-student interaction. "[Teachers] must constantly adjust their methods and materials on the basis of their identification of the local needs of their students" (Tarone & Yule, 1989, p.3). Most Japanese college students are enrolled in one or more classes with names like 'Freshman English,' 'English Conversation,' or 'Oral English' regardless of whether or not on their own free will since these courses are usually compulsory. Although rarely voiced, students do have expectations concerning learning objectives. As Wenden (1990, p.169) states, "... adult learners bring expectations to their language learning based on their previous educational experiences ..." and they are usually optimistic when it comes to attaining actual or perceived improvement in their oral English ability.

A typical first-year student at a Japanese university has had the mandatory six years of English before entering, three years in junior high school and three years in high school. The common resulting phenomenon from the years of studying English in the Japanese educational system is that the students generally have a solid background of fundamental English grammar and a basic vocabulary for beginning level students. A common deficiency, however, is that the students have little or no communicative ability. To strengthen this inadequacy emerges as an identifiable student need and it is then the challenge to the teacher to successfully address it.

It becomes particularly challenging to determine if, and then to what degree, the teacher has met the students' needs. There are unlimited options of rather objectively determining a student's degree of success or improvement in a host of language skills. For example, tests can identify strengths and weaknesses in the areas of grammar, vocabulary, reading comprehension, sentence structuring, listening comprehension, and error identification to name a few. But it becomes much more difficult to objectively assess level and improvement in the realm of oral skills.

Paradoxically, these skills are the ones that can be the source of the most concern and anxiety for students. Cultural inhibitions and individual circumstances aside, who has not heard a story or two about the Japanese student who had a high score on an institutionally recognized test, but could not utter a word when confronted with a seemingly ordinary question from a native speaker? One source of finding out whether or not students' needs have been addressed and determining if the students have indeed improved their spoken English that is often overlooked, especially in a traditional Japanese educational setting, is the students themselves. This paper will define an attempt to use the students as a source of teacher and course evaluation through the means of a questionnaire. The results will then be analyzed and discussed.

A total of 252 students took part in the teacher and course evaluation project. The students were enrolled in a four-year women's university. The course met year-round, that is to say that there were approximately fifteen ninety-minute class meetings in both the first and the second semesters. It was after completion of their final exam of the second semester that students were asked to complete the evaluation. The form consisted of ten questions and a space for additional comments (see Appendix A). With respect to the students' level of expression, the form was written in both Japanese and English. The students' course name and section number were requested, but students were told not to write their names on the evaluation form with the hope that anonymity would increase the objectivity of their responses. Students were asked to rate the teacher and course on a scale with five gradients: 'Poor,' 'Below Average,' 'Average,' 'Above Average,' and 'Excellent'.

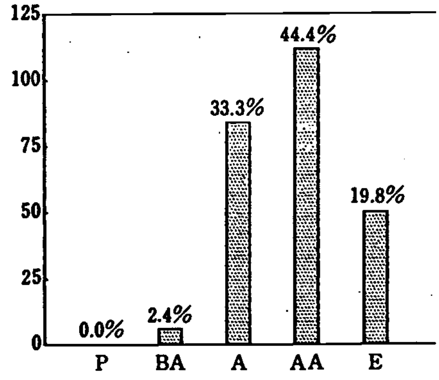
Teacher and course evaluations have intrinsic merit amounting to more than just a popularity contest. A teacher who voluntarily subjects him/herself to the potentially subjective opinions of language learners makes a few inherent statements about his/her teaching philosophy. For example, the teacher believes that the results themselves are worth the time and effort involved to tabulate, translate, read, analyze, and interpret. In addition, the results are worth the risk that there might possibly be some critical information that could be a source of ego-bashing for a sensitive instructor. By utilizing a teacher and course evaluation, a teacher makes the statement that improving the potential of the class and subsequently the level of the student's English is more important than the aforementioned risks and efforts. There is always the possibility that the instructor will discover a previously unthought of aspect of his/her classes, lessons, or techniques and gain insight into the student's learning. After all, it is impossible for an instructor to see his/her teaching form the eyes of each and every one of the students. Evaluations give a teacher access to student perspective, and are at least one way to help a teacher become more aware of student need identification and student self-assessment of improved oral English skills. Furthermore, sometimes the results can be enlightening, revealing, positive, and even humorous.

Statistical Analysis and Interpretation

With more than 250 students answering ten questions, a total of slightly more than 2500 responses were generated. This amount of data automatically lends itself to number comparison. The five options given to students to choose from are represented as follows: P = Poor, BA = Below Average, A = Average, AA = Above Average, E = Excellent. Of course the question of arbitrariness can be posed. In other words, what exactly do 'Poor,' 'Below Average,' 'Average,' 'Above Average,' and 'Excellent' mean?
Teacher Evaluation/6 'Poor' in terms of what? 'Average' in terms of what? 'Excellent' in terms of what? The validity of the gradients on the rating system can only be interpreted through the individual life experience of the students. Being naturally subjective, possible influencing factors include all or some of the following: pre-course expectations, previous English learning experiences in junior high school, high school, college and other private schools, previous teachers at those institutions, post-course self-assessment of improved (or regressed) language skills, and whether or not the student felt that the teacher adequately identified and satisfied his/her needs. It goes without saying that outside factors could effect the tone and attitude a student has when filling out the evaluation. If the student were 'having a bad day,' or were simply tired before completing the teacher and course evaluation, then that could naturally be reflected in the results. However, when taking into consideration the sheer numbers generated by the surveys, it can be interpreted that the numbers depict an accurate overview of the course, incorporating a reasonable margin of error of ±10%.

Each question of the survey is worded to address a fundamental pedagogical question concerning teachers and courses. Questions 1, 2, 4, 5, 6, and 7 (see Appendix A) either directly or indirectly have to do with the evaluation of the instructor. Questions 3 and 8 are closely connected to the course and curriculum. Questions 9 and 10 ask the student to do some introspection and focus on him/herself. For example, if the teacher were concerned about the fairness of the method of testing and evaluation of students, question number 2 (see Appendix A) could provide some insight. Several individual numerical results are interesting to observe. The results of question number 1, 'In general, how would you rate this instructor as a teacher?' are shown in the following graph (Fig. 1):

Figure 1 - Distribution of responses to question #1



On this question, 162 responses or 60.3% fell into the 'Above Average' and 'Excellent' categories. Combining those results with the results of the 'Average' category, there are 256 responses. In other words, a full 97.6% of the students surveyed thought that the teacher was at least average or better. Only a mere 6 responses of 252 or 2.4% felt that the teacher was deficient. The conclusion drawn is that the teacher in general had a successful year in meeting the students' needs for an instructor during the courses included in the survey. The corresponding graphs and percentages for all of the questions on the survey are listed in Appendix B. Refer to these for a complete breakdown of the survey results.

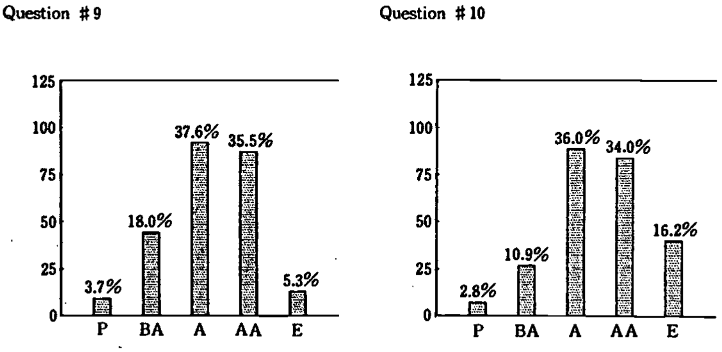
Question number 9, 'How much improvement in English do you think you made from this course?' and question number 10, 'How would you rate your own study habits and the effort you made in this course?' were the two questions on the survey that required students to do some self-assessment. The results of these two questions are shown in the Figure 2.

A glance at the breakdown of these two questions shows that the results from the 'Below Average' and 'Poor' categories had higher percentages than the results from the corresponding categories from all of the other questions on the survey. Accordingly, the 'Excellent' category had the lowest percentages of all questions. It is interesting to note the correlating distribution of responses. It is difficult to argue the importance of out-of-classroom participation when it comes to making progress in language learning. Rubin (1987, p.17) asserts, "It is essential for students to be able to control their own learning process so that they can learn outside the classroom once they are on their own." It seems that students who rated their improvement in English as minimal similarly rated their own effort.

Written Highlights

Perhaps the most useful section of the teacher and course evaluation form was the final part where students were asked to write comments about the ten questions or offer suggestions for improving the course (see Appendix A). Most students chose to write their students in Japanese, they were then translated to English. Although it may have been tempting for students to completely disregard the written comments section, it was encouraging to note that almost 75% of the students who filled out the evaluation took the time to write down their thoughts, and in some cases completely filled a page. The original written comments that the students made are perhaps even more subjective than the pre-determined ten questions. However, they also probably depict a more accurate picture of what the students' needs

Figure 2 - Distribution of responses to questions #9 and #10



really are. Although improvements for the teacher and the course were specifically solicited, it was comforting to find out that in the end not all of the comments were negative nor critical, and in fact, most were positive. Several common reoccurring themes appeared in the almost 200 pages of comments. Some of the highlights are illustrated.

The teacher in these courses made it a point to institute an 'Only English' policy in the classroom. The rules of the policy were explained on the very first day of class, and students were reminded and encouraged to use only English throughout the year in order to reap the pedagogical benefits. At the end of the year more than 40 students made written comments praising the practice of total immersion in English during class time. Some typical comments were as follows:

> "The teacher spoke only English in class, which was first very difficult for me. However, I later realized that my listening skills had been greatly improved."
> "I think what was great about this class was that students were not allowed to speak in Japanese. In other words, we were forced to speak in English to learn that we actually can speak in English."

It was refreshing to see so many students gleaned awareness of their improved language learning through just one teacher-instituted policy. Of course not all students agreed with it as shown below.

> "I understand the importance of total immersion in English. However, this class was difficult for most of us, due to the lack of grammatical knowledge and vocabulary on our part. I'd expect the instructor to use Japanese when it's necessary. I was sometimes lost when he explained only in English."

All in all, there were literally almost one thousand comments about the teacher and courses. Naturally, they ranged from the mundane:

> "This class was fun"

to the insightful:

> "At first I hated this class because I wasn't used to expressing myself. However, I now realize that it worked positively for me because I'm more aware of the importance of having my own opinion and expressing it."

> "I learned that I am the one who has to take responsibility for improving my English. I have to make the effort."

One might not expect an eighteen-year-old first-year university student to have such an awareness about his/her language learning.

Reading through all of the written comments proved to be very informative. Many comments concerned individual class activities, the ones that the students felt the most and the least beneficial. There were suggestions on how to increase class speaking time and efficiency. Gaining insight to how the students perceive a teacher, a technique, a lesson, a class, a course, a curriculum, or an administration is always a challenge for an instructor. Written feedback from the students may be one of the best ways to see a teacher's lesson as the students do.

Conclusion

Teacher and course evaluations can be a useful tool for a teacher willing to gain insight into the way his/her lessons are being perceived. The students themselves are the best resource from which to elicit commentary or criticism. There are an unlimited number of teacher, course, and curriculum related topics that can arise. The use of the target language or the student's native language in the classroom is just an example. Others include homework issues, testing, lesson organization, teaching techniques, lesson activities, cultural topics and differences, class size, and seating arrangements to name a few, and the list goes on. Of course there are risks involved, there is always the possibility that students will criticize and have negatives comments about an instructor or course. However, the risk is worth taking especially when a teacher stands to gain an increased awareness about his/her classes. A problem or alteration is more easily rectified after it has been identified.

Another quote from one of the teacher and course evaluations read as Teacher Evaluation/ 14 follows:

> "I suspect that you won't change your teaching style."

This seems to be a simple comment. One interpretation for the teacher is that student feedback can have as much or as little impact on teaching and course structure as the teacher sees fit. It is impossible to please all of the students all of the time or as Gaies (1983, p.191) puts it, "What was surprising to me was how different students reacted to what went on in the classroom period." Within the realm of a classroom there will be conflicting opinions on any given issue. It is up to the teacher's judgement to filter feedback before deciding whether or not to implement change. Holec, (1987, p.150) sums it up as follows, "The management of evaluation involves both passing types of judgement and using the results yielded as a basis for keeping or modifying the learning program."

References

Gaies, S.(1983).Learner feedback: An exploratory study of its role in the second language classroom.In H.Seliger & M.Long (Eds.), *Classroom oriented research in second language acquisitio*, p.191. Cambridge,MA: Newbury House.

Holec,H.(1987). The learner as manager: Managing learning or managing to learn? In A.Wenden & J.Rubin (Eds.),*Learner strategies in language learning* (p.150). Englewood Cliffs,NJ: Prentice Hall International.

Rubin,J. (1987). Learner strategies: Theoretical assumptions,research history and typology.In A.Wenden & J.Rubin (Eds.), *Learner strategies in language learning* p.17. Englewood Cliffs,NJ: Prentice Hall International.

Tarone, E., & Yule, G. (1989). *Focus on the language learner*. Oxford: Oxford University Press.

Wenden,A.(1990). Helping language learners think about learning.In R.Rossner & R.Bolitho (Eds.). *Currents of change in English language teaching*. p. 169. Oxford: Oxford University Press.

APPENDIX A—Teacher and Course Evaluation

教員および科目評価
TEACHER/COURSE EVALUATION

＊学籍番号・氏名は記入しないこと
Do NOT write your name or student number on this paper.

教員名
Teacher ______________________

科目名
Course ______________________

学科名
Section ______________________

記入例 ：このカフェテリアの食事をどう評価しますか。
Example ：How would you rate the food in the cafeteria ?

○	○	☒	○	○
良くない	あまり良くない	普通	まあ良い	大変良い
Poor	Below Average	Average	Above Average	Excellent

1． この教員を教師としてどう評価しますか。
In general, how would you rate this instructor as a teacher ?

○	○	○	○	○
良くない	あまり良くない	普通	まあ良い	大変良い
Poor	Below Average	Average	Above Average	Excellent

2． この科目の学生の評価の仕方，テストの仕方をどう思いますか。
How would you rate the evaluation and testing of students in this course ?

○	○	○	○	○
良くない	あまり良くない	普通	まあ良い	大変良い
Poor	Below Average	Average	Above Average	Excellent

3． この科目を構成の点でどう評価しますか。目標や方向性ははっきりしていますか。
How would you rate this course in terms of its organization, clarity of objectives and directions ?

○	○	○	○	○
良くない	あまり良くない	普通	まあ良い	大変良い
Poor	Below Average	Average	Above Average	Excellent

4． 1回の講義の構成をどう評価しますか。目標や方向性ははっきりしていますか。
How would you rate the lessons in terms of their organization, clarity of objectives and directions.

○	○	○	○	○
良くない	あまり良くない	普通	まあ良い	大変良い
Poor	Below Average	Average	Above Average	Excellent

5． この科目は興味をそそられ，熱中させられ，刺激のあるものですか。
How would you rate the interest, enthusiasm, and stimulation the instructor brings to this course ?

○	○	○	○	○
良くない	あまり良くない	普通	まあ良い	大変良い
Poor	Below Average	Average	Above Average	Excellent

6． この教員の講義をわかりやすくするためのプレゼンテーションの仕方や説明の能力をどう評価しますか。
How would you rate the instructor's manner of presentation and ability to explain in a clear and understandable fashion ?

○	○	○	○	○
良くない	あまり良くない	普通	まあ良い	大変良い
Poor	Below Average	Average	Above Average	Excellent

7． この教員の学生に対する態度はどうですか。（学生への配慮，関心，敬意）
How would you rate the instructor's attitude toward students (concern, interest, respect) ?

○	○	○	○	○
良くない	あまり良くない	普通	まあ良い	大変良い
Poor	Below Average	Average	Above Average	Excellent

8． 全体のカリキュラムから考えた場合の，この科目の重要性をどう評価しますか。
How would you rate the importance of this course in terms of its suitability in the overall student curriculum ?

○	○	○	○	○
良くない	あまり良くない	普通	まあ良い	大変良い
Poor	Below Average	Average	Above Average	Excellent

9． この科目によってあなたはどのくらい英語が上達したと思いますか。
How much improvement in English do you think you made from this course ?

○	○	○	○	○
良くない	あまり良くない	普通	まあ良い	大変良い
Poor	Below Average	Average	Above Average	Excellent

10． あなたがこの科目のために勉強したり努力したことをどう評価しますか。
How would you rate your own study habits and the effort you made in this course ?

○	○	○	○	○
良くない	あまり良くない	普通	まあ良い	大変良い
Poor	Below Average	Average	Above Average	Excellent

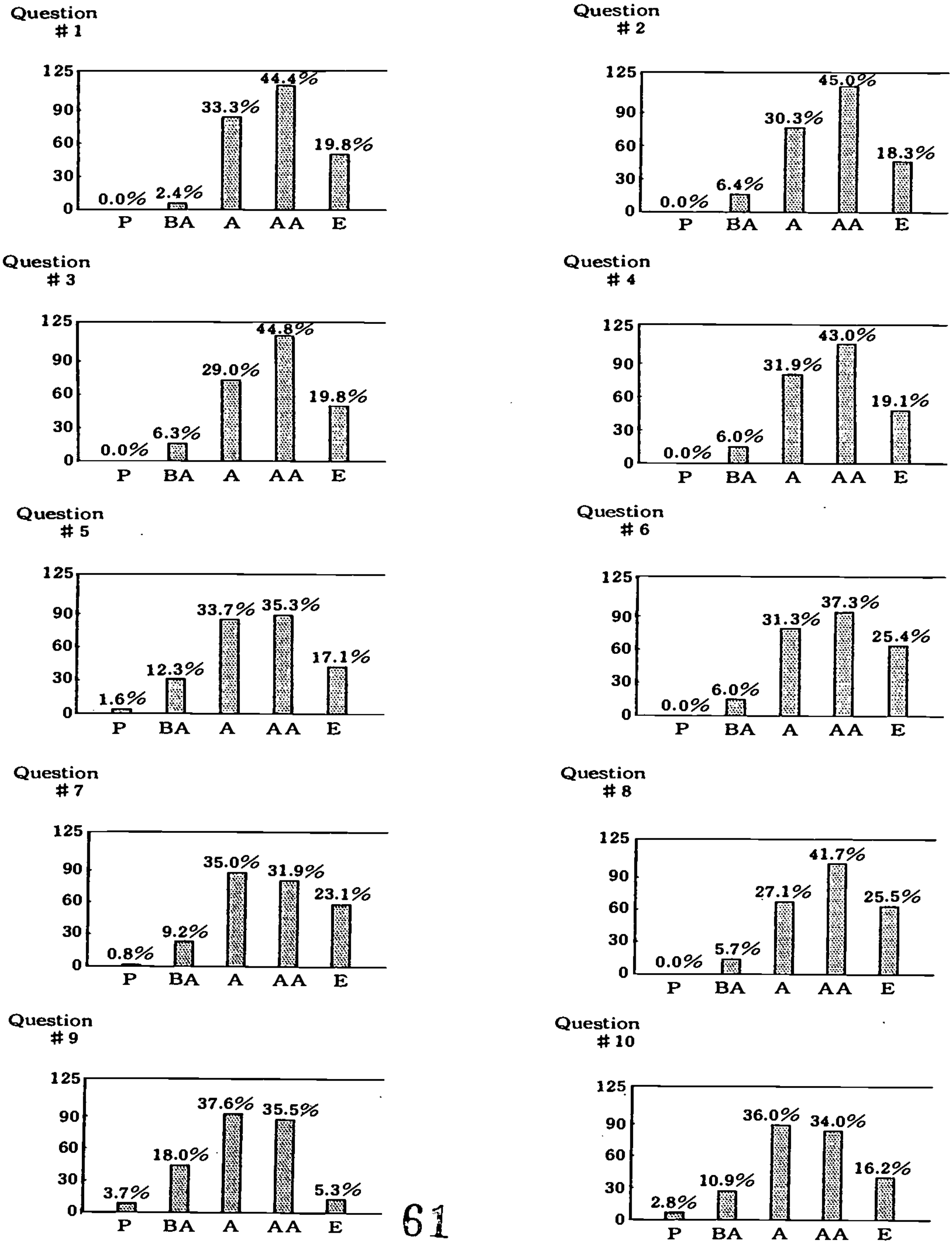
コメント：この科目をもっと良くするための意見，提案などがあれば，上記のことを含め，下の余白に自由に書いて下さい。必要なら他の用紙も使って下さい。

Comments: Please feel free to comment about any of the preceding areas, or offer suggestions you might have for improving this course. Use an additional piece of paper if you need more space.

Appendix B – Statistical and Graphic Representation of Teacher and Course Evaluations.

Chart 1 – Total Responses of 252 Evaluations from 5 Classes.

Key : P = Poor 良くない
BA = Below Average あまり良くない
A = Average 普通
AA = Above Average まあ良い
E = Excellent 大変良い



Question #1

Question #2

Question #3

Question #4

Question #5

Question #6

Question #7

Question #8

Question #9

Question #10

ERIC REPRODUCTION RELEASE

I. Document Identification: ISBN 4-9900370-1-6 (Language teaching; conference proceedings)

Title: On JALT 95: Curriculum and Evaluation
Proceedings of the 22nd Annual JALT International Conference on Language Teaching/Learning

Author: Gene van Troyer, Steve Cornwell, Hiromi Morikawa (eds.)

Corporate Source: Japan Association for Language Teaching (JALT)

Publication Date: July, 1996

II. Reproduction Release: (check one)

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in Resources in Education (RIE) are usually made available to users in microfiche, reproduced in paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. If permission is granted to reproduce the identified document, please check one of the following options and sign the release form.

[XX] Level 1 - Permitting microfiche, paper copy, electronic, and optical media reproduction.

[ ] Level 2 - Permitting reproduction in other than paper copy.

Sign Here: "I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Gene van Troyer Position: JALT President, proceedings editor
Gene van Troyer

Printed Name: Gene van Troyer

Organization: Japan Association for Language Teaching

Address: JALT Central Office
Urban Edge Bldg. 5th FL
1-37-9 Taito, Taito-ku
Tokyo 110, JAPAN

Telephone No: 03-3837-1630; (fax) -1631

Date: October 20, 1996

III. Document Availability Information (from Non-ERIC Source):

Complete if permission to reproduce is not granted to ERIC, or if you want ERIC to cite availability of this document from another source.

Publisher/Distributor: JALT (Japan Association for Language Teaching)

Address: (See above)

Price per copy: ¥2500 (US$25.00) Quantity price: Standard bookseller discount

IV. Referral of ERIC to Copyright/Reproduction Rights Holder:

If the right to grant reproduction release is held by someone other than the addressee, please complete the following:

Name: *None*

Address:

V. Attach this form to the document being submitted and send or fax to:

Acquisitions Coordinator
ERIC/CLL
1118 22nd Street, NW
Washington, DC 20037
FAX: 202-659-5641
TEL: 202-429-9292