Standardised Assessment of a General English Course
(英語授業における評価方法の平準化について)

Mark Brierley & Andrea Orlandini

Key terms: performance, competence, rating scale, communicative testing, standardization

The comprehensive English programme (総合英語) began with the bold aim of providing over 1000 students with a coordinated curriculum, comprising equivalent quantity and quality of material, and leading to fair assessment. This paper will describe a case study examining attitudes to this programme with specific reference to communicative language testing. Thirty teachers from diverse backgrounds were involved in teaching the programme; the eighteen members of staff who responded to a questionnaire were part-time and full-time teachers, native and non-native speakers of English with a variety of formal teaching qualifications and experience.

Our research centred on this questionnaire, which asked respondents to state their opinions concerning the assessment context, communicative strategies and procedures for assessing the course. The results suggest that *Sogo Eigo* is a positive step, supported by respondents; this support is substantial as it is based on agreement with the principles of the course, and the principles of communicative language testing. Some terms popularly associated with communicative language testing, on the other hand, received less support. This ambivalence is less significant, and could be caused by the ambiguous nature of the terms.

1. Why Assess Students?

It has been estimated that around 60% of university lecturers' non-research time is spent on assessment. In some cases, grades can make a difference to the careers and lives of students and assessment is clearly an important part of the teacher's job that is taken very seriously. Before considering what should be assessed and how it should be assessed, we should first consider the reason for assessing in the first place.

Shohamy (1992: 7) points out that "it is common practice in many educational systems to utilize tests to affect and drive learning; in these systems, tests have become powerful devices capable of changing and prescribing the behaviour of those who are affected by their results—administrators, teachers, and students. Central agencies and decision makers, aware of the authoritative power of tests, have used tests to impose curricula, textbooks, and teaching methods. Thus tests that were originally intended to provide information for making decisions and judgements about individuals and programs have become efficient devices for promoting and enforcing teaching and learning. This phenomenon, often referred to as the washback effect[1], illustrates the power of tests to affect test takers' lives".

Undoubtedly, an imperative of *Sogo Eigo* was to integrate assessment with teaching. Thus many test items were incorporated into the program. On the surface, the results of the questionnaire indicate some ambivalence regarding the role of testing. Almost half of respondents (47%) agree that "There are too many assessment items for *Sogo Eigo*", which suggests that testing should play a more limited role in driving learning. However, in contrast 78% percent of respondents claim that testing should motivate students—in other words, be a driving force in their learning.

Why this contradiction? It is a general puzzle that we shall investigate. The next section provides an overview of other opinions, and other paradoxical results, which shall also be referred to throughout the course of the paper, as we explore the nature of communicative language testing, and its consequences in this university.

1.1 Teachers' opinions and Questionnaire Methodology

A questionnaire concerning assessment was given out to teachers of the comprehensive English programme. A total of 18 teachers responded, representing 60% of those teaching the course. This included nine native English speakers and nine non-native speakers, with an average of 17 years' teaching experience. Over two thirds of teachers (70%) had taught outside Japan, and various combinations of teaching experience ranged from elementary schools to adult language schools, including private language schools and business classes. Half the teachers had received formal assessment training, for example from modules of MA and MEd courses or in training to be assessors or interviewers for tests such as Cambridge ESOL. Around half had received no formal training in assessment.

The questionnaire was written in English. While not the native language of all respondents, it was felt that, since all were English teaching professionals, their level of English proficiency would not invalidate the results. The questionnaire was trialled on 4 teachers, and misunderstood items were modified for clarity. The majority of questions presented statements and asked whether respondents strongly agreed, agreed, neither agreed nor disagreed, disagreed or strongly disagreed. The variety of opinions held by teachers is shown by the fact that, out of 45 questions, there was no disagreement on only seven (percentage of responses agreeing in parentheses—the remainder neither agreed nor disagreed):

Statements eliciting a broad consensus	% agreeing
"Assessment should measure whether, or how much, the students have met the goals of the course."	(94%)
"Assessment should motivate students."	(78%)
"We should assess the process (whether students are performing activities and completing tasks) as well as the product (the final report, test or presentation)."	(89%)
"Graded samples of students' work would make it easier to assess writing."	(94%)
"I feel confident in assessing the general writing ability of my students."	(78%)
"Students should keep some record of what books they have read (for example, book reports, book list, etc)."	(94%)
"We should try to assess the writing ability of our students"	(88%)

Table 1 Statements eliciting a broad consensus

The following statements characterize a variety of responses, from a qualified consensus to no consensus:

Statements eliciting a qualified consensus, and no consensus	Agreed	Neither	Disagreed
"Teachers should assess ability (how good the students are at English)."	82%	12%	6%
"To assess communicative ability, teachers need more training."	67%	28%	6%
"To assess communicative ability, teachers need more assessment instruments."	65%	24%	12%
"It is easy for me to tell when students have used translation software."	67%	28%	6%
"It is easy for me to tell when students have plagiarised work."	72%	11%	17%
"The amount students read [in the course's extensive reading programme] should be part of the assessment."	78%	11%	11%
There are too many assessment items for *Sogo Eigo*."	47%	24%	29%
"Self assessment should be taken into consideration in students' grades."	39%	28%	33%
"Grammar should be taught implicitly."	40%	33%	27%
"We should teach specific grammar points in class."	39%	28%	33%

Table 2 Statements eliciting a variety of responses

In many areas there was no consensus. Only half the respondents agreed "We should try to assess the communicative ability of our students", with one person disagreeing, the remainder expressing ambivalence. While a slight majority agreed that there should be regular vocabulary tests, 50% disagreed that there should be regular grammar tests. Both kinds of tests were generally less popular with native speakers than with non-native speakers, and native speakers were generally more in favour of communicative testing, and including both effort and ability in students' assessment.

In answering numerical questions, teachers stated that 25% of the grade should come from final tests (standard deviation 13), 23% of the grade should come from Extensive Reading (standard deviation 11), and students should read 14 books per semester (standard deviation 3.7).

1.2 Course Goals

The case of English language education at Japanese universities presents an anomaly. The traditional role of universities is for experts to provide classes in their specialised subjects. Language skills and communicative ability are far from specialised, particularly if they are to be taught by scores of teachers to thousands of students, as in the comprehensive English course which we will consider in this paper. In most university courses, the lecturer is best qualified to determine the curriculum, teach the course, and assess the students. This can lead to policies stating that assessments should test whether students have learnt what has been taught. Clearly this is appropriate for specialised subjects; however, where a wide range of teachers are following their own curricula under the same course name, such a policy is problematic as it may lead neither to standardised curricula nor standardised assessment.

The following requirements for English programs were specified by the University in 2006:[2]

1. To provide students with the same quality and quantity of studies—in order to fulfil this, common teaching materials should be prepared
2. To provide unified assessment criteria for fair grading
3. To provide materials which are somewhat related to students' major—general English should be a preparatory stage to be linked to their major in the future. (This request came particularly from the science faculties.)

The university's mission statement calls for its students to be "skilled communicators", as well as making a commitment for the university to be "the main driving force for international exchange between the people of Shinshu and the global community". With the status of English as a global language, either one of these provides a mandate for the teaching of English for communication, which is identified as a goal, along with presentation skills, global and international awareness and problem solving skills in the School for General Education's Foreign Language Curriculum (2006).

The Comprehensive English course (*Sogo Eigo*) was developed in response to these requirements. A key aim of the course was to motivate students to embrace the English language as a communication tool that will open the door to a wide range of information, institutions and people throughout their studies, careers and lives. To this end, a range of tasks that combine the four skills build upon and activate linguistic knowledge from the six previous years of compulsory English education.

Rather than teaching knowledge, the course is based on a functional notional syllabus, aiming to develop the skills and practical abilities of students, as they work together and alone to complete each stage of various projects, within specified topic areas. These topics expand from the students in concentric circles, beginning with their surroundings and immediate lives, going on to students around the world, then to future careers and job opportunities and finally to the connection between their majors, the world and the future. In this way, the topics progress from the students' place in the class to their place in the university, then within society as a whole and finally within the global community. Acquisition in grammar and vocabulary come about through working towards goals and learning communicatively. The course is divided into 4 components:

Semester	Curriculum objectives
First	Students look at themselves, where they are studying and what is happening in their lives. Students find out about their classmates through a survey, and work together to produce a newspaper about cultural and social topics that interest them.
Second	Students investigate courses around the world that that they could join. Students then research historical figures and events within their majors, and finally perform debates on issues that put their lives and majors into a global context.
Third	As students begin studying their majors in earnest, jobs and careers are considered, beginning by exploring students' opinions, beliefs and hopes for the future. Next they are called upon to find international organisations where they could work, and finally simulate going through the process of applying and being interviewed for a job.
Fourth	Students look at their majors in the media. First, students will investigate current trends and future developments in their majors by accessing the English-language media. Focusing on factual information, they will write an essay and create questions to evaluate the knowledge of their classmates on the topic.

Table 3 The Comprehensive English course curriculum objectives

Stephen Krashen's Input Hypothesis is taken as a theoretical model of language acquisition within the course, with particular emphasis on the affective filter[3]. In accord with this, extensive reading has been incorporated into the course, and each semester students are expected to read around 100,000 words, at a suitable level for their ability.

Assessment enhances the main goals in *Sogo Eigo* by encouraging students to embrace English as a communicative tool. A wide range of assessment procedures drive communication in the following ways (Shohamy 1992: 14-15):

1. Self-assessment is based on the understanding that learning can be improved when the learner has a central role in planning and evaluation and that effective assessment allows the learner to participate actively in the learning process.

2. Observation as a form of assessment in Sogo Eigo does not obstruct learning, as learners are unaware of being observed. Thus testing does not become invasive.

3. Simulations designed to approximate real-world language use promote authentic language interaction.

4. Projects, interviews, and assignments as assessment procedures encourage a broader range of language use and skills.

5. Peer assessment is less anxiety-inducing than being assessed by a teacher, and can increase the motivation to learn.

1.3 Assessment beyond Proficiency

Another university policy that must be taken into account is a guideline stating that students must attend two thirds of classes in order to qualify for a grade. The university is not solely interested in the proficiency of its students, but is also concerned with their attitude. Such concern is reflected in responses to our questionnaire in that 78% of the teachers who replied agreed that "Students' grades should depend both on their ability (How good their English is) and their effort (How much work they do)". On average, teachers stated that 56% of student assessment should be based on ability and 44% on effort (standard deviation 13). This is also implied by teachers' overwhelming agreement on assessing the process as well as the product of learning (89%). Two reasons can be postulated why students should be given credit for the effort they put in: first, other things being equal, it seems likely that more effort will lead to more learning. There is no clear link between effort and performance, but for many teachers this was an issue (56% agreed with the statement, "I would feel more confident about assessing effort, if there was a clear relationship between effort and ability"). Secondly, and perhaps more importantly, students should be given credit for working hard and trying. There is, however, no theory or research we are aware of which backs either of these points.

Of greater concern, perhaps, is the lack of reliable ways to measure effort. Such assessment runs the risk of being subjective and relying in part on the relationship between the teacher and the student. However, it may be that the ability to build such relationships, while beyond the theories of language assessment, is a virtue to be valued for students' future careers or lives. Such questions are beyond the scope of this paper, which will go on to investigate the nature of communicative ability.

2. What should we assess?

The history of language testing can be divided into three distinct periods: the pre-scientific, the psychometric-structuralist, and most recently the psycholinguist-sociolinguist. The most recent, the psycholinguistic-sociolinguistic appearing in the late 1970's and early 1980's, is popularly known as communicative language testing.

It may be helpful at this point to clarify the terms "assessment" and "testing". The word "test" conjures an image of rows of students sitting behind desks writing silently on identical pieces of paper, while "assessment" implies other, less formal evaluations of student performance, which may include submitted assignments, or an impression the teacher builds up over several lessons based on the raising of hands, answers to questions, and even reaction to what is said. A classical definition may imply that a test is something that takes place after the teaching has finished, although if we take into account the notion of washback, tests clearly have a very important role in the learning process. Brown (2004: 4) uses concentric circles to describe assessment as a subset of teaching, and testing as a subset of assessment, "test" defining something formal, while "assessment" covers a wide range of activities by students themselves, students' peers and the teacher. In this paper, the two terms are used interchangeably, as they often are in literature in the field where it seems that use of the word "assessment" is increasing, perhaps corresponding to a broadening of assessment practice from conventional tests.

2.1 Communicative testing

Morrow (cited in Fulcher, 2000; 489) claimed that there were specific criteria that could be used to tell if a test was communicative. Despite the surrounding controversies, his work has left a legacy in language testing, in view of the wide acceptance of communicative testing. It is now generally believed that communicative testing should involve performance, authentic tasks, and behaviour-based outcomes. Fulcher (2000: 489–93) groups Morrow's list of seven criteria defining communicative testing into these three general categories, into which assessment in the *Sogo Eigo* program fit. This paper limits the discussion to one category, the notion of performance, and excludes discussion on authenticity and real life outcomes. The notion of performance could be broken down into three dimensions:

a. test takers should actually have to produce language

In other words, communicative tests involve performance.

b. there will be actual (face-to-face) interaction which involves not only the modification of expression and content but also an amalgam of receptive and productive skills

In the majority of cases, as Morrow points out (1981: 16), language use is based on an interaction. Even cases such as letter writing may be considered a weak form of interaction in that they involve an addressee, whose expectation will be taken into account by the writer.

c. language use is often in real-time

The processing of unpredictable data in real time is a vital aspect of using language (ibid: 16).

Judging performance is a vital aspect of communicative testing; however, the relation between ability and performance can often be misunderstood. To illustrate, the questionnaire clearly confirms that the majority of respondents support the position that ability should be tested (82%). However, communicative testing which involves testing vital components of ability in both speaking and writing did not receive such support. The next section will explain the relationship between performance and ability.

2.2 A linguistic model of communicative testing

Much work has been done in Canale (1983: 5–14) on the notion of performance, which has undergone much discussion and development in relation to how we should comprehend it in linguistic terms. Before outlining these developments, it should be noted that they were incorporated as a unifying framework in the *Sogo Eigo* curriculum. The distinction between communicative competence and performance—actual communication—is an important one for the *Sogo Eigo* curriculum. Actual communication is the realization of communicative competence, which could be defined into the three broad areas of sociolinguistic, discourse, and strategic skills:

1. Performance involving interaction requires sociolinguistic competence—this type of competency concerns the extent to which utterances are produced and understood appropriately in different contexts depending on contextual factors such as status of participants, purposes of the interaction, and norms or conventions of interaction.
2. Producing language means making discourse—this type of competency concerns the mastery of how to combine grammatical forms and meanings to achieve a unified spoken or written text in different genres.
3. Processing data in real time requires strategic competence—this component concerns the mastery of verbal and non-verbal communication strategies that may be called into action for two main reasons:

   a. to compensate for breakdowns in communication due to limiting conditions in actual communication, or insufficient competence in one or more other areas of communicative competence

   b. to enhance the effectiveness of communication

Support for communicative testing was implied in the questionnaire results, which showed that most teachers thought more assessment instruments should be provided (65%) and training in communicative testing should be made available (67%). The next section outlines standard methods for assessing communicative competence.

2.3 Standard methods for assessing communicative competence

According to Spolsky (1985: 180), three main methods are used in assessing language performance. Observation involves a recognition of the relationship between performance and ability. Self-assessment involves a recognition of the relationship between process learning and goal learning. Task-based assessment involves a recognition of the relationship between goals and learning outcomes.

A: observation e.g. recording, analysis, and judgement as in the COLT observation scheme[4].

Only samples are necessary to indicate whether students have attained a certain level of proficiency. A sample of performance could yield information about communicative competence indirectly through observation. Observing the performance two or three times would improve reliability by confirming the student's proficiency in the competency area. Communicative competence is an essential prerequisite for actual communication or performance, but is reflected in performance indirectly through observable behaviour.

In principle, observation in communicative testing is supported by the questionnaire insofar as 89% of respondents suggest that the steps involved in the learning process leading to the goal should be assessed. The observable performance of students depends upon the communicative competencies defined in Canale (1982). For example, if learners maintain their interaction with clarity and cohesion, we may infer that they are using discourse strategies, and have a high level of discourse competence. If the interaction is not cohesive, we may infer a lower level of discourse competence. In relation to sociolinguistic competence, if students seem to be interacting appropriately as befits the instructional context, we may infer a higher level of sociolinguistic competence, and vice versa. Rating scales generally operationalize levels of performance and proficiency, enabling systematic observation schemes. This will be discussed in more detail in section 3.

B: introspection or self report e.g. self-assessment sheets/or checklists.

Spolsky (1985: 181) points out that self report is considered satisfactory in situations where no special degree of accuracy is required. If the goal should be to teach strategic competence, then performance in self-assessment tasks could yield data that contributes to information about students' level of strategic competence. More importantly, there are many taxonomies categorizing communicative, affective, meta-cognitive and cognitive learning strategies which teachers can use to assess strategic competence.

The questionnaire also indicates a possible case for using self-assessment, even though it is not supported directly. Self-assessment is clearly a controversial area, only 34% indicating it should be used in grades. On the questionnaire, one teacher commented: "Self assessment should not be included in their final grade. It is for their own sake." Another identified a conflict between self-assessment and standardisation thus: "...one problem with standardising assessment is that it makes it difficult at best for assessment to be handed over to the students—i.e. with students themselves deciding what will be assessed and how—or by getting students themselves to set and mark exam questions." According to Gipps (cited in Oscarson 1997: 175) effective and relevant learning is best achieved if the student is actively engaged in all phases of the learning process, according to which the learner's own reflection on and creative restructuring of already acquired concepts, understanding, and points of learning play a crucial role in the building of new knowledge. As 94% of respondents agreed that students should be assessed on whether the goals of the course have been met, if the development of strategic competence is a goal of the course, then this suggests that self-assessment could at some future point be included in grades.

C: formal elicitation of performance e.g. completion of tasks and activities

The abilities underlying performance (discourse, strategic, and socio-linguistic competence) may be assessed on the basis of whether students could achieve their task outcomes. Did they exchange opinions? Did they solve a problem? Did they cooperate to solve a task? Did they close an information gap? Report sheets could objectively measure whether discourse outcomes are achieved, if realization of the tasks depends on using those language skills.

An issue with the completion of tasks is use of the target language within the classroom. If students are using their first language rather than the target language, completion of a task may not necessarily have relied upon proficiencies in the language we wish to assess. Therefore, assessment should involve a variety of methods and measuring instruments that can allow the assessor to triangulate data to make more reliable judgements about students' proficiency.

While almost all respondents (94%) indicated that assessment should measure whether the goals of the course are achieved, it should be noted that the term *goal* maybe interpreted in a variety of ways. If the goals of *Sogo Eigo* should continue to be communicative, and we are to measure whether these goals have been met, then the standard methods of assessment outlined above need to be continuously refined and developed. In approaching this objective in *Sogo Eigo*, we have identified rating scales as a key instrument.

3. Standardization and rating scales

The Comprehensive English program has sought to use rating scales for many of its assessments. The issues that emerged as a result of using rating scales in *Sogo Eigo* 1 will be addressed and discussed in relation to the main theme of standardization in section 4, and section 5 will suggest future directions for improving standardization practices. The following section will prepare the way with a detailed examination of the nature of rating scales.

3.1 What are rating scales?

A widely-used way of operationalising language proficiency is in the form of language proficiency rating scales containing descriptions of different levels of ability. In many cases, they have been developed in response to increasing demands on educational institutions to report test scores in ways that make it clear what people can do in the test language. There is a very real need for standardisation as departments of Japanese Universities seek accreditation, for example with JABEE, the Japan Accreditation Board for Engineering Education. This is echoed in policy statements, outlined above, for quality and quantity of teaching to be equal, and assessment to be fair.

Although rating scales describe performance, most rating scales claim to be measures of *underlying* competence. Performance is thus interpreted as an indirect indicator of ability. In other words, the level of students' strategic, sociolinguistic, and discourse competence can be ascertained by comparing their observable performance to levels on a rating scale. The questionnaire is implicitly in favour of using rating scales, 94% of respondents agreeing that graded samples of students work would make grading easier. Graded samples are normally used in conjunction with rating scales as in the case of the Cambridge tests illustrated below.

Rating scales consist of a series of descriptions of stages or ranges of language behaviour in one or more language skill areas along some kind of continuum of increasing ability which often ranges from zero to native-like. The level definitions typically describe the kinds of tasks and texts that learners can handle at different ability levels and the degree of skill with which they can achieve various communicative goals. For example, the speaking tests of Cambridge ESOL's PET, FCE and CAE assessments specify four analytical categories: Grammar and Vocabulary, Discourse Management, Pronunciation and Interactive Communication. Descriptors are given for 5.0, 3.0 and 1.0; 3.0 representing adequacy, 5.0 representing top of the range and 1.0 indicating inadequate performance. The assessor gives a score within 0.5 marks for each category, so there are nine mark bands: 1.0, 1.5, 2.0 ... 4.0, 4.5, 5.0. A second assessor, who also acts as interlocutor in the test, gives one score on a global scale (French, 2003: 8).

3.2 How should rating scales be administered in Sogo Eigo?

Ratings are assigned to candidates by eliciting a sample of language performance under test conditions. With speaking and writing, this is usually done by having trained assessors compare learners' observed performance with the descriptions on the scale. The content and organization of level descriptors have been developed in a variety of ways by different institutions to meet various needs.

The question remains as to how the content and organization of rating scales should be developed and adapted for *Sogo Eigo*. An attempt to identify the needs of Shinshu University's General English Department might start with a recognition that testing can become a negative force if the content and organization of rating scales are inappropriate. The examples below represent 5 possible rating schemes that could be considered for *Sogo Eigo*:

- The English Speaking Union yardstick provides general indicators of overall communicative ability.
- The Australian Second Language Proficiency Rating Scale gives relatively detailed descriptions of particular features of language use typical of each level.
- The American Council on the Teaching of Foreign Languages refer to global tasks/functions,

context, content, accuracy and text type as criteria for describing levels.

- The Royal Society of Arts Certificates in Communicative Skills in English refer to complexity, range, speed, flexibility and independence as criteria for describing levels.
- Cambridge ESOL has three tests at a suitable level for students at this university. FCE, IELTS and BEC, which use rating scales to assess candidates' performance in paired interviews.

3.3 Advantages of rating scales

There are a number of clear advantages and disadvantages provided by rating scales (Brindley 2003, p60-61):

- They enable reporting to external audiences in the form of qualitative performance descriptions rather than just scores.
- They are focused on what people can do rather than on what they know. Because the descriptors relate to language performance, rating scales provide a direct link back into the curriculum, thus enabling the teacher to focus on proficiency goals.
- The levels and descriptors can provide a common language for teachers to describe individuals and classes and recommend suitable activities or teaching strategies
- Scales and their associated elicitation procedures are easier to understand and have a more obvious connection to the curriculum than standardized tests. They are hence more likely to be acceptable to practitioners.
- Because they explicitly describe different levels of ability, rating scales can be used to sensitize language learners to the gap between their current level and their desired goals.

3.4 Disadvantages of rating scales

Brindley also cites a number of sources indicating several disadvantages underlying the use of rating scales (2003: 61):

- The lower levels tend to be negatively expressed in some scales, which is de-motivating to some learners and can fail to acknowledge good performance on lower level tasks.
- A number of writers are cited (Lumley and McNamara 1995; North 1993) to indicate that scales are inherently unreliable because assessors interpret and apply the rating criteria differently. Although training clearly improves coordination between the scores given by different assessors, ESOL reckoning on correlations of 0.9 between their oral examiners, research has established that in spite of intensive training, significant and ongoing differences in rating persist.
- Alderson's (1991) critique of the IELT speaking scales is cited to point out that the descriptors used in many general rating scales are very vague and impressionistic. As a result, it is difficult to specify relative degrees of mastery of a particular skill with sufficient precision to distinguish clearly between different levels.
- The fixed and hierarchically ordered level descriptions fail to take into account the variability which is inherent in second language use. For example, one level descriptor may contain different skills which develop at different rates.
- Bachman (1990) is cited to point out that the level descriptions are highly context dependent

and thus do not allow generalizations about underlying ability.

- Lee and Musumeci (1988) is cited to raise the point of construct validity. It is difficult to find any explicit information on how the descriptors which figure in many of the high-profile scales were actually arrived at, partly through issues of test security, so their basis in research remains unclear. Consequently, the criteria which constitute the levels have usually not been independently verified, and the relationship between task and level tends to be based on intuition and experience rather than research.

If careful attention is paid to the curriculum needs of the institution when selecting, adapting and applying rating scales, these disadvantages will be reduced and the above advantages fostered. It may therefore be necessary to select scales which most directly complement the goals of the course, and the specific institutional needs of the General English Department. Before considering future directions, it is necessary to consider what problems were encountered in *Sogo Eigo* 1, and why.

4. Assessment for Sogo Eigo I

Assessment for the first semester of *Sogo Eigo* was designed with regard to the principles of general education, the goals of the course and the tasks created for the course. The nature of general education seeks to include students with a range of abilities, and a principle aim of the curriculum was to motivate students. As failure is highly de-motivating, criteria were set so that assessments could be passed by students for whom English was not a speciality.

The desire to motivate the students, and a commitment to place emphasis on input and tasks rather than the study of grammar and vocabulary led assessment away from conventional paper tests, and indeed, after a pre-test in the first lesson to assess the initial proficiency of the students, there were no paper tests until the last lesson of the second semester. This is not entirely sanctioned by teachers, with 67% supporting end of year tests, and 50% supporting end of term tests. On average, teachers recommended that 25% of the grade should come from such tests (standard deviation 13). Combined with ambivalence towards grammar and vocabulary tests, this shows that there is clearly some mandate for including communicative testing.

Many language courses in academic institutions are obliged to meet the criteria of an external test, and when the exam results are published, it is clear whether the students, the teacher and the course have met their goals. In the case of *Sogo Eigo*, tasks, activities and teaching materials were first designed to meet the goals, and then assessment was integrated where possible. A core principle of assessment—that assessment should be repeated as many times, and in as many ways as possible—dictated that teachers should seek to assess a variety of tasks and activities. From a long list of everything that could be assessed, the following short list was decided upon, according to ease of application and relevance. With 47% agreeing that "There are too many assessment items for *Sogo Eigo*" and 29% disagreeing, it is clearly difficult to meet a balance between the needs of maintaining reliable assessment, and reducing teachers' workload.

4.1 Assessment items

As shown in table 1, Assessment for *Sogo Eigo* I was split into nine items, such as performance on a written assignment or the completion of a worksheet. Where possible, teachers were asked to give two different scores for an assessment, for example, one for the quantity of work produced and the other for its quality. Each *Sogo Eigo* class was taught by two teachers, with 90 minute classes on two different days each week. These teachers are referred to as 'A' and 'B'. Generally speaking, the A teachers were more concerned with providing language input to the students, while the B teachers were responsible for organising activities in which students produced linguistic output.

Item	Task	Medium	Criteria
A (1) Self Introduction	writing of a short self introduction	Essay	Rating scale (Accuracy, clarity)
B(1) Survey Results and Discussion	Project in which groups worked together to conduct a class survey	Worksheet	Rating scales (Quantity of work) and (clarity and organisation)
B(3) Speaking Performance I		Observation	Rating scale (Participation)
B(4) Self-Assessment	Answer questions about studying habits	Worksheet	Full marks for submission of paper
B(5) Extensive Reading Activity (W9B) (Paper)	Group activities based on books read	Worksheet	Rating Scale (Creativity, structure, content)
A(2) DVD Comments as a film director	Essay based on video material	Essay	Rating Scales (Creativity) and (Organisation)
A(4) Self-Assessment	Answer questions about studying habits	Worksheet	Full marks for submission of paper
A(5) Speaking Performance in class	Throughout semester	Observation	Rating Scale (Participation)
B(6) Speaking Performance II	Throughout semester	Observation	Rating Scale (Participation)
B(7) Speaking Performance III	Throughout semester	Observation	Rating Scales (Communicative ability)
B(8) Extensive Reading	Extensive reading	Book reports	Number submitted by students

Table 4: Assessments for Sogo Eigo 1

Throughout the semester, each teacher gave scores for each assessment. These scores were added together within the Blackboard e-learning platform to give a percentage for each student. *Sogo Eigo* 1 used only assessor-oriented scales to rate the performance of students e.g.

A(1) Self Introduction. The task was the writing of a short self-introduction.

5—The student has written the whole page with fairly limited errors. The paragraph is clearly written, and one can see good points here and there.

4—Though there are a few more errors in writing than the above, one can see good points in some places, and in general it is clearly understood. The paragraph filled more than two thirds of the page.

3—It is difficult to comprehend in some places, and some basic grammatical errors can be found. However, in general one can have a clear picture of the student's self-introduction. More than a half of the page is written.

2—Many grammatical errors are found, and structures are quite distorted. One can understand what the student means only with a guess. There are a few sentences one can understand.

1—The paragraph is not comprehensible, however, one can guess that the student has written about him/herself.

0—The student did not submit the worksheet

4.2 Performance of assessment instruments

The main variation in grades for the first semester of the Comprehensive English Course seems to be the way that the assessment criteria have been interpreted. All teachers are experienced in assessment, and spend a lot of time and effort on assessing fairly and accurately, as previous guidelines have specified. However, all teachers have developed their own systems, their own scales and their own pass-marks, based on widely different teaching experience, in terms of length, breadth, subjects taught, and nationalities, ages and locations of students, and all have received different formal education in language assessment practice, some having received none. Therefore, some teachers pay more attention to attitude and participation, some to creativity and imagination, some to grammar and vocabulary, and each teacher has a different definition of what constitutes adequate, or exemplary performance.

As a consequence, when we compare two different scores, we cannot be certain whether they are different because the students are performing differently, because the teaching has been different, or because the assessment has been different. All we can do is compare scores with other scores, compare scores between the two teachers teaching each class, and look for differences.

It was possible to make three comparisons between the scores given by each teacher to the students: results for speaking participation (A5 versus B6), with identical assessment criteria; results for all writing quality assessments (A1 and A2 versus B1 and B5), each with different criteria; and total scores given by each teacher, except for extensive reading. It was found that of the three comparisons, the total scores showed the highest correlation, with an average of 0.30 over 35 pairs of teachers. The highest correlation between two teachers was 0.76, while one pair of teachers showed negative correlation (-0.33). The average of the correlations on writing assessments was 0.24, with a range from 0.53 to -0.17. Where teachers were using the same criteria (see below) to assess the same students on speaking participation, the correlation averaged at 0.09, ranging from 0.59 to -0.26.

Clearly, many teachers agree broadly on the students' performance, and some pairs of teachers agreed to a high correlation. However, there were systematic differences in either their assessment of, or the students' performance in speaking activities.

It should be noted that we cannot expect two teachers to agree completely on their assessment of students. Indeed a very high correlation would more likely suggest that teachers were being influenced by each others' results. Let us remember that if we seek to assess communicative performance, it is impossible to reach 100% reliability. All language assessment steers between the two beacons of reliability and construct validity. Construct validity dictates that tests should test what they seek to test; on the other hand if a test will give the same results at a different time, on a different day, or under different circumstances, then it is considered to be reliable. Rather than meeting both ideals, there is usually a compromise between them, or a choice: either a test can be reproduced reliably and fairly, or it accurately represents some kind of authentic ability.[5] The University of Cambridge Local Examinations Syndicate introduced a compulsory oral component in its tests as far back as 1913, and in 1945 John Roach published a paper looking into the many issues involved in oral testing, entitled *Some problems of Oral Examinations in Modern Languages* (Taylor, 2003: 2–3). These problems have not yet been completely resolved, as demonstrated in Alderson's critique of IELTS (cited in Brindley, 2003; 61). Ultimately, communication cannot be assessed objectively as communication itself is not an objective activity. If we wish to assess communication then we must accept lower reliability than we may get on, say the TOEIC test[6]. However, if we do not assess it, we take a much greater risk of failing to take communication seriously as an educational goal.

4.3 Interpretation of Rating Scales

Although it is impossible to independently verify most of the assessments given by teachers, we can infer that the interpretation of rating scales and assessment criteria has varied from teacher to teacher. Another general problem was that there was little differentiation. In some classes, twenty or thirty percent of students were separated by only one or two percentage points. In other words, a slight change in a cut-off grade could change many students grades from A to B. The correlation between teachers demonstrates that there is some agreement between teachers as to which students are 'better'. However, the actual scores that they give may be so different that in one class almost all students would receive A grades, while in another, almost none would. If the institution was simply interested in whether students passed or failed, this would not be critical. However, the university requires differentiation between A, B and C grades, and assessment procedures should work towards the ideal that a student who received an A in one class would have received an A from any other teacher.

The evidence all suggests that there is variation between teachers in the way assessment criteria have been interpreted. The following section will address directions that could be taken to reduce this variation.

5. Proposals for Sogo Eigo Assessment

Experience in *Sogo Eigo* 1 and the questionnaire results suggest that more work is needed in developing performance criteria for rating scales. The results of the questionnaire confirm that *Sogo Eigo* is in principle a step in the right direction with emphasis on communication supported both by the institution's policy and its teachers. However, there remains a potential for variability in the interpretation of rating scales. The following proposals are therefore directed specifically at the way rating scales are designed, implemented, and managed.

5.1 Criterion-referenced testing versus norm-referenced testing

All assessments are either criterion-referenced or norm-referenced: criteria-referenced if they are measuring candidates against fixed criteria, for example as we may hope in the administration of a driving test; norm-referenced if they are comparing candidates against each other, as may be appropriate in an entrance exam for a university where a fixed number of places need to be filled by the best candidates. According to Brindley (1991: 143), classroom criterion-referenced assessment is often concerned with assessing learners' attainment on a scale of ability which represents varying degrees of mastery, but is not necessarily linked to a cut-off score such as a pass mark. The questionnaire makes a clear case for continuing the use of rating scales. The fact that the majority of respondents are in favour of criterion-referenced assessment as opposed to norm referenced assessment is suggested by the claim by 94% of the respondents that the goals of the course should be measured by assessment. In other words, rating scales are undoubtedly supported by the questionnaire results.

If rating scales are to be used with minimal disadvantages and maximal advantages as outlined in sections 3.4 and 3.5, then consensus must be built among teachers on their construction, meaning and application.

5.2 Specifying the communicative criteria that will be tested

Designing specifications for testing is a formidable undertaking and could play a significant role in maximizing advantages and minimizing disadvantages in the use of rating scales. Descriptors and criteria should give users a clear view of standards by which performance is judged based on the goals of the course. Alderson et al. (1995: 20–21) specifies a number of important requirements in designing specifications:

- typical examples of performance at each level
- a description of a what a candidate achieving a given grade can achieve in the real world
- examples provided of candidates' performances on previous tests
- a description of how the criteria used to assess those performances apply to the examples
- a description of what test preparation would be appropriate prior to the examination.

5.3 Implementation and practical considerations

Alderson et al. (1995: 111–113) suggest that appropriate members of staff decide, for example, what level descriptors are suitable for the University departments, or decide with samples what the level descriptors mean. Bachman (1996: 50–52) suggests that workshops could involve issues such as standardizing characteristics of setting, test rubrics, input, and expected responses. In the case of Shinshu University's General English Department, the standardising committee's roles may be:

- Reviewing the performance of the assessment items
- Revising criteria
- Removing or modifying unsuccessful items and introducing new items
- Revising rating scales
- Providing graded models to demonstrate what is meant by rating scales
- Planning implementation of assessment
- Planning evaluation of assessment
- Reviewing once again, and repeating the process for the following semester.

Once criteria have been decided, much work is needed to apply them to the students in the classroom. This will be difficult without a consensus among teachers upon the assessment goals, criteria and wording of rating scales, and will be impossible without agreement on what those scales mean in relation to actual samples of student output. Key questions therefore include the relationship between competence and performance, the nature of communicative competence and methods for formatting and managing assessment instruments. For example, should external examining bodies be relied upon to test communicative ability, should teachers conduct oral interviews, or should teachers observe student performance in the course of interactive classroom activities? If the last of these options is to be chosen, a number of classroom management issues need to be resolved, including patrolling schemes allowing teachers to systematically observe every student.

Another practical consideration that must be taken into account to facilitate the implementation of the assessment is the collection of samples of oral performance, which are required for three reasons: to set standards, to provide graded models for teachers to base their assessment on, and to validate assessments of different teachers.

In addition, recordings could be the basis of research into students' performance at communicative activities, leading to further improvement in the curriculum. We believe that the answer may lie in recent solid state technology, which means that there is no longer a need for the purchase and use of cassettes, and recordings may be quickly and easily transferred, copied and stored electronically. Such resources would be sensitive to budget constraints; at the time of going to press, such equipment can be purchased for under 4,000 yen although it should be noted that prices of electronic equipment are constantly falling.

The potential benefits of on-line grammar testing must also be considered. If made part of the course grade, teachers would very likely appreciate a lower assessment workload, allowing more time to concentrate on communicative testing, and the preparation of materials and classes. Thus, another area where technology may assist assessment is by increasing the amount of computer-based testing for the course, requiring students to complete assessments that could be automatically graded. Teachers seem sceptical of computer testing, only 22% agreeing that "There should be on-line tests that all students have to do", with 33% disagreeing and 44% ambivalent. Only slightly fewer are sceptical of using computers for teaching grammar, 35% agreeing that "Specific grammar points would be better taught by online activities" with 29% disagreeing and 35% ambivalent. Before any implementation, further investigation is necessary into the reasons for such scepticism, and whether increased familiarity and reduced assessment workload would put computer based testing in a more favourable light.

We believe that standardised assessment is likely to be made easier by increasing teachers' involvement in the process of managing assessment, from as early as possible. Not only would this provide valuable insight into the practicalities of assessment, but would begin earlier the lengthy process of coordinating different assessors to implement assessment items in the same way, and assess candidates fairly. For this purpose, regular meetings could be scheduled in which teachers may work together, not only discussing the goals of the course and principles of assessment, but also working on the fine detail and comparing their assessments of specific students.

5.4 Proposals for future research

The process of applying assessment criteria could benefit from action research. Nunan (2001: 197) has called for a change of emphasis in research literature from discussions of methodology to investigations of practicalities within specific educational contexts, arguing that such research is every bit as valid as traditional academic research. The development of communicative testing within the university could provide valuable research in an area that is of wide relevance within the field of applied linguistics as well as many teaching situations in Japan where communicative policies of schools and governments do not correspond to the examination or test syllabi.

The *sogo eigo* curriculum, and communicative assessment in particular present many issues begging for active research, ranging from the development and implementation of communicative assessment instruments to the application of computer based testing. Issues for active research could also include investigations into teacher attitudes towards standarisation in curricula, the teaching of vocabulary and collocation, and the kind of English we should be teaching, for example where our students will use the language, for what purpose and with whom. We eagerly await such research and its publication.

6 Conclusion

In conclusion, the university's policies clearly support communicative assessment, and judging by the results of our questionnaire, the English teachers support its principles. The Comprehensive English course has created an environment in which this can happen and is already happening. To further promote communicative testing, consensus must be built among teachers, to achieve common objectives and methods of such assessment. Moreover, context-sensitive research is necessary, as Nunan (2001) argues, to find solutions that meet the specific institutional needs of the General English Department.

Acknowledgements

The authors wish to thank all colleagues who have participated in the Comprehensive English course and who completed the questionnaire which is the basis of this research. They would also like to give particular thanks to Kenichi Ariji, Sean Mehmet and David Ruzicka, for understanding, encouragement, and constructive criticism.

Notes

[1] See, for example, Hughes, A. (1989). Testing for Language Teachers. Cambridge: Cambridge University Press, p1 or Brown, H.D. (2004). Language Assessment – Principles and Classroom Practice. White Plains NY: Longman, p28.

[2] Based on the document 'Strategic Planning for curriculum development in general education in 2006' (17 November, 2003). And surveys conducted in 2002 and 2003 by visits to each Faculty.

[3] See Krashen, S. (1992). Fundamentals of Language Acquisition. Chicago, IL: SRA/McGraw-Hill, pp. 1–11.

[4] For details of the COLT (Communication Orientation of Language Teaching) observation scheme, see Genesee, F. and J. Upshur (1996). "Classroom-based evaluation in second language education". Cambridge: Cambridge University Press (Chapter 5, pages 76-97), p78.

[5] The Atlantic presents a divide between emphasis on reliability in America, where features that cannot be assessed reliably are not assessed, to emphasis on validity in Europe, where attempts are made to assess desired skills, even if this cannot be done accurately. This is echoed in the joke that

"there is a country in Europe where multiple-choice tests are illegal" (attributed to Sigfried Hulzer), multiple-choice items representing everything that is desirable in terms of reliability and derisible in terms of construct validity for communicative testing.

[6] See Buck, G. (2001). Assessing Listening. Cambridge: Cambridge University Press, p210–216, for an analysis of the TOEIC test with regard to the listening construct.

References:

Alderson, J. C., C. Clapham and D. Wall. (1995). Language Test construction and evaluation. Cambridge: Cambridge University Press.

Bachman, L.F. and A. S. Palmer. (1996). Language testing in practice. Oxford: Oxford University Press.

Brindley, G.. (2003). "Language Testing and Evaluation." Unpublished Lecture Notes. Macquarie University.

Brindley, G.. (1991). "Defining Language Ability: The Criteria For Criteria" In S. Anivan (ed.). Developments in Language Testing. Singapore: Regional Language Centre.

Canale, M. (1983). "From communicative competence to communicative language pedagogy." In J.C. Richards and R.W. Schmidt (eds.) Language and Communication. London: Longman.

Fulcher, G.. (2000). The 'communicative' legacy in language testing. System, 28, 483-497.

French, A. (2003). "The development of a set of assessment criteria for Speaking Tests." In Research Notes, 13, September 2003. Cambridge: Cambridge University Press. http://www.CambridgeESOL.org/researchnotes, accessed 25th January, 2007.

Morrow, K. (1981). "Communicative language testing: evolution or revolution?" In J. C. Alderson and A. Hughes (eds.). Issues in Language Testing. ELT Documents 111. London: The British Council.

Nunan, D. (2001). "Action Research in Language Education." In D.R. Hall and A. Hewings (eds.) Innovation in English Language Teaching. London: Routledge. 197–207.

Obana, Y (2005). 共通教育英語への希望―教員と学生へのアンケート結果報告 [New English curricula in general education—a report on the questionnaires] Shinshu Journal of Educational Research, Centre for Development of Higher Education, Shinshu University, 1, 109-120.

Oscarson, M. (1997). "Self-assessment of foreign and second language proficiency." In Clapham, C. and D. Corson (eds.), Encyclopaedia of Language and Education, 7: Language Testing and Assessment. Dordrecht, the Netherlands: Kluwer academic Publishers, 175-187.

Shinshu University (2006). 共通教育外国語カリキュラム [General Education Foreign Language Curriculum]. 高等教育機構設置準備室外国語教育部門 Department of foreign Language Teaching, School of General Education, Shinshu University.

Shohamy, E. (1992). "New modes of assessment: The connection between testing and learning". In E. Shohamy and A. R. Walton (eds.) Language Assessment for Feedback: Testing and other Strategies. Dubuque, Iowa: Kendall/Hunt Publishing Company.

Spolsky, B. (1985). 'What does it mean to know how to use a language? An essay on the theoretical basis of language testing' in Language Testing, 2, 2.

Taylor, L. (2003). "The Cambridge Approach to Speaking Assessment." In Research Notes 13, September 2003.

(Associate Professor, School of General Education, Shinshu University)
(Lecturer, School of General Education, Shinshu University)

23/ Apr. /2007 Accepted