Toyama University of International Studies (TUINS) is a small, private university that was founded in 1990 and has an enrollment of about 1,000 students. This paper will consider reliability and validity factors in a representative English language test that was part of the entrance examination for the university in 1999. In order to understand the nature of the English test and the analysis put forward, the context in which the entrance examination was taken by prospective students will be outlined first.

Not all students needed to take the entrance examination or an English test in order to be admitted to TUINS. About 30% of the 193 students that entered the university in 1999 were accepted from recommended high schools on the basis of an interview in Japanese, an essay written in Japanese, and their academic record. Thus, they were not required to take any written entrance examination at all. In fact, less than half of the enrolled first-year students took an English language test.

The entrance examinations were given at three different times in that year. The first time, in January, all candidates had to take the English test and one other test from among Japanese language, Japanese history, and world history. Candidates sitting for the entrance examination the second time it was given chose just one test to take from among the four above and mathematics. The rationale was to attract more applicants at that time by allowing them to take a test in the subject area in which they thought they would do best. The third time it was offered, in March, shortly before the start of the school year in mid-April, candidates simply chose either the English test or a Japanese language test. At that time relatively few students sit for the entrance exam and it is a practical matter to offer a minimum number of tests.

The English test that will be analyzed in this paper is the one that was offered in March, 1999. 22 Japanese high school seniors took the test and 21 were accepted for admission by the university, of which nine went on to enroll (41% of the test takers). From among all those who took one of the three English tests, 73% (189) of the 258 candidates received scores above the cut-off for admission and 48% (91) of them actually enrolled (35% of all the test takers). The cut-off scores were determined primarily by the enrollment capacity of 100 (set by the Ministry of Education) for each of the two departments of the university, international studies and social studies, and the number of places still open at the time of the administration of each of the entrance examinations. Thus, although the mean scores (out of 100 points) of the three English tests were the following (in order of their administration): 47, 51, and 56, the cut-off scores varied considerably: 29 and 32, 40 and 61, and
51 and 34, for the international studies department and the social studies department respectively, for each of the three tests in order. Some consideration was given as well to the absolute scores themselves in determining the cut-offs.

While these may seem to be complicated procedures in the admissions process, they are representative of those followed by universities in Japan, particularly those that are private, and pointedly so in recent years when from 1992 to the present, the population of 18-year-olds nationwide, the nearly total pool of university entrants in Japan, has declined more than 25%, and often the paramount consideration of an institution is its survival.

The purpose of the TUINS English test, with the other tests offered in the entrance examination, is to screen applicants for general academic achievement in their high school studies on the basis of their performance on the test. The English test is a part of the entrance examination for the university as a whole. It is part of an assessment of candidates for all courses they may take at the university. The test is intended to be similar in content and form to the English portion of entrance examinations of other universities and the general university entrance examination used by public universities, preparation for which has been the core and the driving force behind students entire high school education. At the same time, the English test offered by TUINS is believed to convey a more communicative use of the language than other such tests and thus to impress students and teachers at high schools with the manner in which English is taught at TUINS. There is, however, no English language or literature major at the university. Rather, students must fulfill a required number of English course credits, which may be taken with Japanese teachers and/or native-speaker teachers.

The full-time staff of English teachers at TUINS in 1999 was made up of three Japanese and three Americans. Every year three teachers are assigned each to make up a draft of one of the tests. Tests of past years are taken as models, although a teacher may include different types of questions. It is agreed, however, that all questions must be multiple choice or other objectively marked types, except for the short writing sample that is included for students to do. The draft of each test is read and taken by all the teachers and then all questions and items are reviewed and critiqued in a plenary meeting of the staff. (I will follow Heaton’s (1990, p. 30) use of the word *item* for very short questions and the word *question* for a number of similar items grouped together.) Questions or items that are believed by consensus to be far too difficult or easy, ambiguous, culture-based and beyond the experience of Japanese students, or inappropriate in any other way, are revised on the spot or resubmitted at a later date by the test drafter. At that time, the same procedure is followed for any newly written questions and/or items until full approval of the test is arrived at.

Following is an overview of the English language test analyzed in this paper. Question I (items 1 - 8) was a task to choose the most appropriate, natural, and grammatically correct response to a conversational question or statement. Question II (items 9 - 13) and Question III (items 14 - 20) were designed to be similar to the TOEFL, section 2 (Structure and Written Expression), parts A and B, sentence completion and error identification tasks, respectively. Question IV (items 21 - 25) and Question V (items 26 - 32) were word meaning recognition tasks in which the same or opposite meanings respectively are sought. Question VI (items 33 - 36) and Question VII (items 37 - 40) were designed to be similar to TOEFL, section 3 (Reading Comprehension), reading and answering questions tasks. Question VIII was a writing task of 40 - 50 words on the given topic. Instructions for all questions were given in Japanese to avoid any lack of understanding as to what to do on the part of the examinees. The time for the entire test was 80 minutes. Two points each were given for items 1-32,
three points each for items 33 - 40, and twelve points for the writing sample. The test was administered at the university by Japanese staff with at least one English teacher present. It was marked by the English teachers, and students were informed simply of having passed or not passed the overall entrance examination.

Reliability is concerned with the construction, administration, and scoring of a test. The underlying reliability questions are: would an assessment produce the same or similar score on two occasions or if given by two assessors? Reliability therefore relates to consistency of pupil performance and consistency in assessing that performance (Gipps, 1994, p. 67). Synchronic reliability refers to comparisons of a test to parallel tests that are different and yet examine the same features. The English test that is evaluated herein may be compared to the other two tests that were given at earlier dates that year. Both were made up of mostly the same question types as the third test. Differences included a sentence word order question on both those tests, and one also had a multiple choice cloze question while the other had a multiple choice phonological discrimination question. The topics for the writing samples, of course, were also different, but the genre of personal narrative account was the same. Given the university entrance examination system in Japan, with the rather strict channeling of students by high school advisors to particular universities according to their test-taking performance, it can be said that students sitting for the TUINS exams were broadly similar in their academic achievement in high school and, more importantly for this analysis, in their overall results on entrance examinations. The similarity of the mean scores for the three TUINS English tests, then, would lead one to assert a good degree of synchronic reliability.

Except for the writing sample, all the questions on the English test are made up of multiple choice type discrete items. There are practical reasons for this, namely, they allow for economy, ease and speed of administration and scoring, and the objectivity ensures high reliability in marking. At the same time, however, the use of a multiple choice format brings with it a number of problems. The multiple choice item is a thoroughly unrealistic measure of language performance. It does not reflect actual language use - there is no real-life situation in which we go around asking or answering multiple choice questions (Underhill, 1982, p. 18). Further, in fact, multiple choice tests may not be any more reliable or valid than similar tests in different formats (Oller, 1979, p. 257).

The preparation of sound multiple choice tests is technically very difficult, particularly item writing. Oller (1979) lists some of the common problems: Selecting inappropriate content for the item,...Including two or more plausible choices among the alternatives,...Leaving unintentional clues about the correct choice,...[And] writing distractors that don’t fit together with the correct choice (p. 245). Indeed, on the last point, Oller notes that the design of distractors to trick the learners into confusing dilemmas is counter-productive. It runs contrary to the very idea of education (p. 256). Multiple choice test preparation certainly is no simple matter. Oller (p. 255) further lists the steps considered necessary for constructing a good test. First, one must have a clear notion of what it is that needs to be tested. Then, item content is selected and item format devised appropriately. After the test items are written, they are evaluated by qualified people for difficulties of vagueness, ambiguity, and lack of clarity, and any weak items rewritten. The items then are pretested and item analysis of the results may involve the examination of item facility indices, item discrimination indices, and response frequency distributions. Pretesting and analysis steps may be recycled until sufficient levels of
reliability and validity are obtained. The item content and format in the TUINS English tests are carefully considered and the editorial work is diligently attended to by the full staff of English language teachers. However, the notion of what precisely is to be tested appears to be assumed rather than rigorously clarified, and pretesting and subsequent technical item analyses are not undertaken.

Some of the difficulties of preparing multiple choice items may be seen in Question I (see the Appendix) of the test considered herein. In item 2, if *pilaf* is not known to the examinee, *(b)* may be considered a plausible answer. Also, *(d)* as the only negative answer, stands out as most likely incorrect. In item 6, *(a)* may be a plausible answer along with the intended correct answer *(b)*, as may be *(b)* and *(c)* in item 7 as alternatives to *(a)*.

The writing sample of each candidate was read by two teachers and scored holistically on a scale from 1 to 12 points. The analytical categories of evaluation taken into consideration included organization, coherence, content, grammar, syntax, and word choice. When the scores of the two readers differed, the mean was taken as the final score. Generally, the scores given were 5 or 6 points, and seldom were the scores of the pairs of readers divergent by more than two points, indicating relatively high inter-marker reliability. There were no checks on intra-marker reliability, however.

An aspect of the topic, *How do you usually spend your New Year’s holiday?*, that was not considered at the stage of test construction caused some difficulty for a number of students. There are several customary Japanese events, activities, and things associated with that time of the year. Not unexpectedly, most students mentioned some of these, but many had difficulty using appropriate words in English to refer to them and this led to some of the writing samples to include Japanese words or some awkward usages of English for these things Japanese. Thus the validity of the writing item was compromised to some degree as well was the reliability of the scoring by the teachers who may not have consistently considered these difficulties.

Face validity simply asks *does this look like a reasonable assessment?*. It is concerned with the reactions of those involved in the procedure. How is it perceived by its users? (Derewianka, 1999, p. 5). By this account, the TUINS English test has a high degree of face validity. As has been pointed out, the content and format accords with the type of English university entrance examination high school students are familiar with and have prepared for years to take through the high school curriculum and the instruction they received.

The test is essentially an achievement test in an easily recognizable format that attempts to reflect what students have learned from their high school study of English. It is used in a normative way to compare the candidates for admission to the university. The test is not intended to have predictive validity, that is, its purpose is not to predict how well students will perform in their English studies at the university. Indeed, in spite of the student-centered, interactive, communicative nature of instruction in the wide range of courses taught by the native-speaker English teachers, there is not at all in the English test a direct measure, nor any items intended as an indirect measure, of speaking and listening skills needed for performance in those courses. Of course, too, the performance of those who enroll cannot be compared with those who scored below the cut-off and were not accepted for admission.

Concurrent validity refers to the degree to which a test correlates with, or gives the same results as, another test of the same type. There is no basis on which to consider such validity as the university has no access to any other test scores of the candidates.

Hughes (1989) provides the following definition of content validity:

A test is said to have content validity if its content constitutes a representative sample of the
language skills, structures, etc. with which it is meant to be concerned....In order to judge whether or not a test has content validity, we need a specification of the skills or structures, etc. that it is meant to cover.... a principled selection of elements for inclusion in the test.... The greater a test’s content validity, the more likely it is to be an accurate measure of what it is supposed to measure (p. 22).

There are a number of problems with the TUINS English test that would seem to reduce its content validity. There is a very limited number of items for each of the skills presumably intended to be measured. Question I is restricted to three functional language items - request, invitation, and asking for advice; a tag question; and four assorted statements and questions. Questions II and III together give a greater number of items for structures, but half of them involve conjunctions. The twelve vocabulary items in Questions IV and V seem an insufficient sample to measure word knowledge, and in the latter question (see the Appendix) the meanings of the words underlined in items 26 and 28-32 could be determined by the contexts of the sentences and thus it may be reading skills that are more being tested. Oller (1979) takes this matter further:

Two nagging questions continue to plague the user of discrete point vocabulary tests. The first is whether such tests really measure (reliably and validly) something other than what is measured by tests that go by different names (e.g., grammar, or pronunciation, not to mention reading comprehension or IQ). The second is whether the kind of knowledge displayed in such tests could not better be demonstrated in tasks that more closely resemble normal uses of language (p. 226).

Construct validity is concerned with the extent to which the assessment task reflects the theoretical assumptions underpinning its construction (Derewianka, 1999, p. 6). The questions on the test, however, are described by the test maker as no more nor less than particular tasks which reflect in content and format what students have studied in high school. The reading comprehension questions, for example, are simply modeled on the TOEFL and identified as tasks of reading and answering questions. They do not necessarily imply a construct of reading such as that of a transaction between the reader and the text in which meaning is constructed in a process that involves one’s background knowledge and experience, or schemata. As Oller (1979) points out: Thus, by naming the task rather than positing some abstract construct we avoid a priori validity commitments - that is, we suspend judgment on the validity questions pending empirical investigation... leaving open the question of what the test is a test of (pp. 234-235). Yet Oller goes on to advise:

The first principle [of item writing], therefore, would be to ask if the material to be included in items in the test is somehow related to the skill, construct, or curriculum that the test is supposed to assess or measure....If a test fails this first evaluation, no matter how elegantly its items are constructed, it cannot be any better than any other ill-conceived test of whatever it is supposed to measure (p. 238).

The writing sample of the English test is an example of a direct measurement of writing ability and as such should not cause us too much concern: even without research we can be fairly confident that we are measuring a distinct and meaningful ability (Hughes, 1989, p. 26). As for the indirect measures of written expression such as those in the form of TOEFL questions (III and IV), Hughes goes on to remark:

Once we try to measure such an ability indirectly, however, we can no longer take for granted what we are doing. We need to look to a theory of writing ability for guidance as to the form an indirect test should take, its content and techniques (p. 26).

This seems to be one of the needs in the construction of the TUINS English test.
As suggested earlier in this paper, the university entrance examination system and the high school curriculum set by the Ministry of Education to support and maintain that system have been widely viewed for many years as one of the fundamental causes for the severe failings of the education system in Japan at the secondary level, and its inadequate preparation of students for the needs and challenges of the new century we have entered. This study is not the place to expound on this issue, yet Gipps (1994, pp. 31-57) offers very cogent insights on this matter from several studies. The principles posited by Madaus (1988) as a result of his survey of the impact of testing on the curriculum, which Gipps cites, all apply with particular force in the case of Japan. Among them are the following:

If important decisions are presumed to be related to test results, then teachers will teach to the test. When test results are the sole, or even partial, arbiter of future educational or life choices, society tends to treat test results as the major goal of schooling rather than as a useful but fallible indication of achievement. A high-stakes test transfers control over the curriculum to the agency which sets or controls the exam (pp. 35-37).

The conclusions from Smith’s (1991a) study regarding the effects of testing on teachers, cited by Gipps, likewise strongly apply to Japan, particularly: The focus on material that the test covers results in a narrowing of possible curriculum, and the reduction of teachers’ ability to adapt, create, or diverge. The received curriculum is increasingly viewed as not amenable to criticism or revision (pp. 50-51). Lastly, the findings from Shepherd’s (1992b) review of research about the effects of high-stakes testing, cited by Gipps, are chillingly apropos to Japan, particularly: The kind of drill-and-practice instruction that tests reinforce is based on outmoded learning theory. Rather than improve learning, it actually denies students opportunities to develop thinking and problem-solving skills (p. 56).

Consequential validity refers to the degree to which assessment procedures enhance or distort the teaching/learning process (Derewianka, 1979, p. 7). Does the TUINS English test, then, add to the negative washback for high school English teaching/learning? Should the TUINS English test include a speaking and/or listening comprehension component(s) in an effort to establish a more communicative approach to teaching/learning at high schools? I believe that the answer to both questions is no. Given the control of the curriculum by the Ministry of Education and the historic, deeply embedded and unyielding entrance examination system, only a major move by the University of Tokyo and other top-ranked public and private universities to alter the process and content of entrance examinations and admission criteria would begin to crack this ruinous system. Although in the near future, a listening component will be included in some university tests, it has not yet been established as a communicative element nor have the consequences of this move for high school curriculum development at all been determined.

Beyond problematic considerations for a speaking component for the English test, namely, time, cost, logistics, and reliability, the face validity of a test with any oral and/or aural elements might be weakened since high school students do not engage in any significant amount of communicative English. They rarely speak English in the classes and they are rarely spoken to in English by the Japanese teachers of English. The impact of now more than 5,000 native English speaking Assistant Language Teachers (ALTs) in junior and senior high schools throughout the country since the mid-1980s seems to be minimal. I am not aware of results of any empirical study that counters this anecdotal impression derived from having taught at post-secondary
level in Japan for 16 years. Given the very limited time they spend in classes, with size typically as many as 40 students, and the fact that most have had no previous teaching experience and are under the direct supervision of the Japanese teacher, nothing more could be reasonably expected.) The predominant teaching methodology of the Japanese teachers of English is yakudoku - a form of grammar/translation (Gorsuch, 1998) which in fact does not very successfully teach grammar or reading, or even effective test-taking skills. (Note that of the more than 165 countries that offer the TOEFL test, Japan places nearly at the bottom in terms of average scores.) Because of the total pervasiveness of the present system, any change in the basic content and format of the TUINS English test would not even have a positive washback in the small Toyama prefectural area, from where about 70% of the school’s enrollment comes. Indeed, such a move by TUINS might lead to the test being perceived as one for a university with lower standards when compared with those given by major universities.

Thus, in spite of some limitations of the TUINS English language test that is part of the university’s entrance examination, and the ethical considerations outlined above, it is perceived to fulfill its purposes adequately, in fact is utilized with considerable fairness, and likely will continue to be constructed and used essentially as it has been. The TUINS English curriculum offers a very wide range of interesting and challenging subject areas and instruction, and in my experience, at least, the overall performance and achievement of students in these courses at the university has been generally at least satisfactory and at times notable.

Selected questions from the TUINS English entrance exam, March 1999:

1. Do you mind if I open the window?
   a) No, go right ahead. b) Yes, please open it. c) No, it wouldn’t. d) Yes, it is fine.

2. This pilaf is pretty good, isn’t it?
   a) Yes, it does. b) Yes, it is very pretty. c) Yes, it is not bad. d) No, it is.

3. Do you have any idea where Mariko is? I haven’t seen her all day.
   a) No, she has been here. b) Not really. She must be at home. c) Neither do I. d) Yes, I believe she has.

4. What is your father like?
   a) Yes, I like him very much. b) He likes baseball and swimming. c) He is rather tall, friendly and funny. d) I like his character and the sound of his voice.


5. Would you like to go to the movies with me?
   a) No, I don't. b) I'd love to, but I already have plans. c) I would go to the movies.
   d) Yes, I would have gone with you.
6. My toothache is getting worse and worse.
   a) So it is. b) You had better go to the dentist. c) That's why I went to the dentist.
   d) Yes, it is a problem for many people.
7. I don't really want to go to that party tonight. What should I do?
   a) If I were you, I just wouldn't go. b) You should buy some snacks and drinks to take with you.
   c) You really ought to go in my car. d) Well, it really is a hard place to find. Maybe going by taxi is best.
8. What is your favorite way to travel?
   a) At night. b) To Egypt, Turkey, and India. c) By train. d) In summer or spring.
   e) For adventure.
9. That book was so dull that I couldn't even finish the first chapter.
   a) significant b) fascinating c) relative d) considerable
10. There has been a slow but steady decline in the number of students interested in studying German.
    a) increase b) assignment c) alteration d) adjustment
11. It seems that no matter how kind and helpful I try to be, I still manage to annoy him.
    a) bother b) please c) encourage d) persuade
12. I feel that the company is really doing a job in trying to foster a sense of company spirit among its workers.
    a) inhibit b) recognize c) promote d) encourage
13. The government is worried that the voters may be shocked by the drastic cuts in spending that it is planning.
    a) orientation b) complete c) complementary d) minor
14. He is so rigid in his opinions that it is all but useless to discuss anything with him.
    a) ancient b) flexible c) conservative d) level
15. I'm quite surprised you like his paintings. I find them rather peculiar.
    a) interesting b) normal c) colorful d) strange