Some Interim Reflections on the Interpretation of English Proficiency Data from the IUJ Graduate School of International Management Admissions Screening of Japanese Students

Richard Smith
English Language Program
International University of Japan

1. Outline of the study

This paper reports the results from a preliminary analysis of data from the admissions screening battery and from first year Grade Point Average (GPA) for Japanese students admitted to study in the MBA program at IUJ. Among other aims, the major aim of the paper is to establish if there is any evidence of a predictive relationship between the scores from the English proficiency screens and first year GPA. A major part of this task is to identify benchmark correlations between TOEFL scores and GPA at comparable English medium graduate institutions and estimate to what extent the observed correlations in other studies are equivalent to those observed in this study.

2. Foreword

This report only lays claim to "interim" status for the following reasons: (a) the subject population of less than one hundred is still small; (b) the data is incomplete; (c) the survey of the literature is confined to ESL/EFL sources. Despite these limitations, it was felt that an interim analysis was justified by the interest in it expressed by members of the Graduate School of International Management (GSIM) and by the opportunity an interim analysis provides for criticism and suggestions for future research. A full report analysis will not be possible until 1995, by which time there should be data for a subject population of well over 100. In view of the fact that the intended audience of this interim report comprises both English language teaching specialists and GSIM content faculty the report will not assume any specialized knowledge of English language tests or of GSIM screening procedures.

3. Introduction

Founded in 1982, IUJ is a two year graduate institution which is one of a small, but growing, band of international English medium universities which are distinguished from most universities in countries where English is the native language by virtue of the fact that a *large majority* of the students are non-native speakers of English. Naturally, this has led both of IUJ's graduate schools, International Management and International Relations, to place even greater than normal emphasis on measuring both the English proficiency and academic suitability of applicants. Both Schools have revised and elaborated their admissions procedures continuously. For Japanese applicants, who are able to attend the entrance examinations in Tokyo, this means that an impressively large amount of data related to a wide range of admissions criteria is collected. GSIM leads the way

with, from 1990, no less than 8 separate measurable criteria, three directly related to English proficiency and some others partly or indirectly related to English proficiency and since 1992 GSIM has employed no fewer than 10 measurable criteria.

Such a wealth of admissions screening data is an ideal realized by the few rather than the many among institutions of higher education and indicates an admirable concern for the academic health of the program and its students. For the ESL/EFL researcher one of the beneficial consequences of this wealth of data is that it is possible not only to investigate the behavior of a battery of English proficiency measures, but also the battery's relationship to the battery of other, more specifically academic, measures. Eventually, it is hoped that this ongoing research will contribute something of value to the already large body of research on admissions screening of non-native speakers of English at English medium institutions of higher education. One valuable property of the GSIM grading system in particular, and of many graduate schools of management in general, is the provision they make against grade inflation and/or distortion by setting restrictions on grade allocations that guarantee something resembling a "normal" distribution of grades.

4. Purpose of the study

This study was motivated by four questions about the GSIM admissions screening battery for Japanese students.

First, what factors inherently constrain the interpretation of the scores from the English proficiency components of the admissions battery?

Second, how should the English proficiency scores be interpreted as a self-contained battery of data? This broad question includes two important specific questions. Can the English proficiency scores be used to predict academic success as defined by GPA ? Do the existing English proficiency admissions requirements discriminate sufficiently to ensure that the academic performance of the Japanese students in a challenging graduate program, which offers keen competition from native speakers and strong non-native speakers from other countries, is not being significantly depressed by a lack of general English proficiency? The answers to these two specific questions are related since if the English proficiency scores of admitted students do correlate highly with later academic performance, it follows that those students who do least well are being held back to a significant degree by lack of English proficiency. These questions can probably never be answered with any great precision because of the difficulty in defining the difference between "strong", "moderate" and "weak" correlations in a field as notoriously complex as education. What can be done, however, is to discover what kind of correlations with later academic performance appear to be the norm at other graduate institutions with roughly similar English proficiency admissions requirements and, after taking account of various moderating variables at these institutions and at IUJ, determine the relationship to the correlational norm of the correlation for the Japanese students in the GSIM.

The third question refers to the relationship between the battery of English proficiency scores and the entire GSIM admissions screening battery. Specifically, in what way can the former be regarded as modifier variables of academic aptitude scores, in this case the Graduate Management Aptitude Test scores?

Lastly, is there any evidence of redundancy in the English proficiency tests used in the admissions battery and, incidentally, in the entire GSIM admissions screening battery?

Since the discussion of the second group of questions occupies the most space in this paper, it should be pointed out to readers outside IUJ that the admissions screening process does not represent the last stage in the assessment of the Japanese students' English proficiency since almost all of them have to attend a ten week Intensive English Program (IEP) prior to matriculation. The focus of this paper is limited to the relationships that can or cannot be established between data collected during the admissions screening process and later academic performance. It is definitely intended, however, that future research will include analysis of the data from the IEP.

5. The subjects of the study

The subject population is confined to Japanese nationals who were actually admitted to GSIM. Later studies will also deal with non-admitted Japanese nationals, but this data exists in somewhat fragmentary form and there has not enough time to collect it.

The faculty members of GSIM are very interested in the performance of the Japanese students as they are natives of the country in which IUJ is located, but non-natives in terms of the language of instruction employed at IUJ. Furthermore, for the moment at least, the Japanese students are in a majority in GSIM.

Another reason for selecting the Japanese students is that they represent a very homogeneous group in terms of nationality, culture, age, sex, working background and so on. This homogeneity reduces the number of variables that can account for variance in academic performance. In addition, the non-residents of Japan, who apply for admission from overseas, are subject to a smaller and slightly different battery of screening criteria.

Data about these Japanese subjects was collected from 1989. The GSIM Program was actually established the previous year, in 1988, but a comprehensive screening process was not put in place until the following year - the starting point for this study. The end point for most aspects of this study is the 1991 intake as the students who entered in 1992 have yet to accumulate a full year of credits towards their degree.

The subject population for the main period, 1989-1991, comprises 61 Japanese nationals, of whom sixty completed the first year of studies and two were women. Other descriptive statistics are not available at the time of writing, though it is known that almost all the students were below the age of forty.

6. The IM School admissions screening criteria for residents of Japan.

Table 1 provides a diagrammatic view of these admissions screening criteria, their score intervals and the date of their entry into the screening battery.

Table 1

Development from 1989 to 1991 of the eight GSIM screening criteria for residents of Japan (Screening criteria directly related to second language English proficiency are in bold).

Screening Criteria		Grading	1989	1990	1991
1. GMAT	Total Verbal Quant	Original Percentile Scores	√	√	√
2. TOEFL	**Total**	**Original Scaled Score**	√	√	√
3. Interview: General Criteria		5 bands		√	√
4. Interview: English Proficiency Criteria		**5 bands**	√	√	√
5. Essay Test: Contents		5 bands		√	√
6. Essay Test: English Proficiency Criteria		**5 bands**	√	√	√
7. Undergraduate Institution		3 bands 1990-1991		√	√
8. Undergraduate Transcript		3 bands 1990-1991		√	√

7. A brief description of the admissions screening criteria

Since the status of the TOEFL test is not entirely clear, it is dealt with at length in Section 8.

The English interview and essay tests (items 4 & 6 in Table 1) have been expressly designed by the English Language Program at IUJ as admissions screening devices which attempt to determine whether or not Japanese candidates have such proficiency in these areas that, with the aid of the pre-matriculation Intensive English Program, they will be prepared for the discussion and writing tasks the students will be required to perform by the MBA Program.

The same interviews and essays are also rated by GSIM faculty according to "general" or "content" criteria (items 3 & 5 in Table 1) that are believed to be helpful in indicating a candidate's suitability for an MBA program.

It is believed that undergraduate records (items 7 & 8 in Table 1) might have some value as performance predictors because, in theory at least, undergraduate behavior is very similar to behavior in graduate courses.

The Graduate Management Admission Test (GMAT) is the only test in the battery which is purposely designed as an academic performance predictor. As the 1991-92 "Guide to the Use of GMAT Scores" (ETS 1991) makes clear on page 5:

"The GMAT is a test of developed abilities intended to provide counselors and admissions officers with one predictor of academic performance in the first year of graduate management school."

This document also goes on to caution that:

"GMAT test scores are but one of a number of sources of information and should be used, whenever possible, in combination with other information "

It should be borne in mind that this test was specifically developed for native speakers of English. The mean scores of non-native speakers of English are well below those of native speakers (Powers 1980).

8. Factors that inherently constrain the interpretation of the English proficiency and other scores

Tests are designed for specific purposes so it follows that interpretations of the data yielded by tests must take acount of their designs.and other charcteristics. This is a fairly straightforward task in the case of most of the tests in the screening battery: they have either been specifically designed as GSIM screening measures or, in the case of the standardized GMAT test, there is a public consensus about its design and use. By contrast, the design and use of the standardized TOEFL test are the subject of a little controversy and, thus, deserve to be discussed at some length.

In addition, no test is perfect. Any interpretation of test data has to begin with an assessment of a test's structural imperfections, such as lack of validity or reliability.

8-1 Design and other inherent constraints on interpretations of TOEFL test data

In recent years the Educational Testing Service (ETS) has gone to great lengths to give the TOEFL test the image of a "pure" language proficiency measure. The statement quoted below is written in bold type in the "TOEFL Test and Score Manual":

"The TOEFL test is a measure of general English proficiency. It is not a test of academic aptitude or of subject matter competence, nor is it a direct test of English speaking or writing ability" (ETS 1992a:18)

It should be noted, however, that the TOEFL test was expressly designed in the 1960's to meet the needs of US colleges and universities which wanted to evaluate the English proficiency of foreign applicants (Oller & Spolsky 1979) with the result that there was some academic bias in the test's design which remains to this day and which has not gone unnoticed by language specialists. When responding to these observations ETS quickly changes track and, as on page 33 of the very same publication cited immediately above, but in regular type, accepts the criticism that the "reading passages tend to be entirely academic in focus" and that the vocabulary item stems in the test "also tend to exhibit primarily an academic content orientation" and claims that this "is consistent with the intended use of the test as a measure of proficiency in English for academic purposes".

Empirical studies of native speakers in the U.S.A. who took the TOEFL test bear out this admission that an academic bias applies to the "Reading and Vocabulary" section of the test. Two studies of college students at the University of Tennessee (Johnson 1977) and of college-bound high school seniors in New Jersey (Clark 1977) who were given the test showed that, although they achieved uniformly high mean scores of 628 and an unscaled 134.5 out of 150 respectively, these native speakers did less well on the reading part test than on the listening and structure part tests. Johnson, for instance, cites a mean reading score of 56.6 out of a maximum 66. Clark

reports a similar performance for his high school seniors on both the reading and the structure test parts.

While it is important not to exaggerate the degree of native speaker fallibility on the TOEFL test - the standard deviation reported by Clark, for example, was only 11.44 -, these findings have two important implications for researchers. One is that there is no compelling *a priori* reason not to investigate the relation between TOEFL scores and academic performance. The other is that researchers are well-advised to consider the TOEFL part scores as well as the total score when conducting such investigations.

8-2. Validity, reliability and calibration: TOEFL

The TOEFL Test is generally regarded as being highly valid in terms of most measurements, with the crucial exception of content validity (personal communication from Perry). The debate over its content validity is related to its perceived academic bias and to its emphasis on evaluating test items in terms only of how they discriminate among candidates. The great virtue of this test is its very high reliability for a language test. Its one-directional Standard Error of Measurement (SEM1) is reported to be 14.1 (ETS 1992a). It has a good theoretical scaled score range from 230 to 677, with intervals of 0,3,7 per ten points (i.e 134 intervals). In practice, however, a score range of 300-677 is "observed" (ETS 1992a:26).

8-3. Validity, reliability and calibration: GMAT

Like the TOEFL test, the GMAT is a relatively reliable test. The SEM1 for both the verbal (GMAT-V) and quantitative (GMAT-Q) sections is 2.4 in a theoretical range of 0-60, with intervals at each of the 60 digits. The total scaled range is 0-600, with intervals at every 10 points and an SEM1 of 24(GMAC 1991:10). By contrast to the TOEFL test, there is little debate about its validity.

8-4 Validity, reliability and calibration: Other tests

No validity or reliability studies have been performed on the interview and essay tests. Because the basic function of the tests is to separate students into a very small number of groups defined as "strong", "average", "weak" etc. and because of the need for speed in evaluating the test performances, the measuring device is broadly calibrated by five bands, with occasional half banding. With the increasing use of half bands the evaluation device is evolving into a ten band instrument and this degree of calibration probably represents the limit for a quickly evaluated test. The standardized "TOEFL Test of Written English", for example, has 11 bands to describe the writing proficiency of the entire non-native population (ETS 1992b:26).

These observations on the essay and interview tests also apply to the other screening procedures listed in Table 1.

Although the uncertainties created by the unknown SEMs are a cause of concern, the most important cause of imprecision inherent in the admissions screening battery is the broadness of the calibration intervals of the non-standardized measures. *The broad score intervals of the non-standardized measures may be inevitable, but they have the important consequence of depressing observed score correlations between them and any other evaluation instrument.*

9. Constraints on the interpretation of the admissions screening scores imposed by a pre-selected population

It is an *a priori* feature of any admissions screening procedure that the candidates who are permitted to enter a serious higher education institution are not representative of a "normally" distributed population. This has important implications for trying to establish relationships between GPA, which is collected from a highly pre-selected set of subjects, and the admissions screening instruments that, to a lesser or greater degree, are calibrated to measure a much broader subject population. As is well known (e.g. Brown 1988, and Farhady & Hatch 1982), observed correlations between scores on different measures are likely to be depressed if the distribution curves of all or any of the measures of the subject population are "abnormal". For the present study two types of abnormality are significant: differential observed score ranges and skewed distribution curves. The Appendix provides a detailed picture of observed ranges of the English proficiency scores, GMAT and first year GPA, and of the distributions of scores within the observed ranges. What follows is a summary of these descriptive statistics.

9-1 Observed Score Ranges

Admitted Japanese nationals' TOEFL scores cover the range 487 to 637, a spread of 150, which represents just under half of the full score range "observed" by TOEFL. Within such a truncated range, TOEFL's SEM 1 of 14.1 represents 9.4 % of the total spread, a significant proportion. The corresponding interview and essay test scores are restricted to two or three of the five grading bands. A similar pattern is also visible in the GMAT test scores. For the GMAT Verbal Test the observed range is 8-35 within a possible range of 0-60, and for the GMAT Quantitative Test the observed range is 29-50 within a possible range of 0-60. The SEM1 for both GMAT-V and GMAT-Q is 2.4 (GMAC 1991:10), which represents 8.9% of the observed GMAT-V spread, and 11.4% of the observed GMAT-Q spread, again a significant proportion. In fact the only distribution that comes close to normal is that for GPA (See Appendix). This is considerably helped by the GSIM grading policy (See the GSIM Handbook for 1992, page 8).

9-2 Skewness

The first year GPA, TOEFL and GMAT-Q scores are relatively unskewed, but those for the English interview and essay tests are negatively skewed and those for the GMAT-V are positively skewed (See Appendix).

The sorts of statistical abnormality identified in this section of the paper are bound to have a somewhat depressive effect on the observed relationships among the various measures. The same sort of picture is also true of the other screening measures not described above. Nonetheless, it should be pointed out that this same picture is generally representative of comparable studies at comparable institutions which use the same or similar screening instruments.

10. Use of admissions screening ratings as predictors of GPA for non-native speakers of English: empirical findings in the literature

This survey of the literature is preliminary and is mostly confined to second or foreign language research sources. The survey of the latter source area has so far

identified *no* predictor studies which are specifically devoted to MBA programs. It is intended in the future to broaden the survey area to include other sources which might be relevant.

10-1 Use of admissions screening ratings as predictors of GPA: Interview and essay screening tests

This survey has yet to discover any meaningful research on the relationship between interview panel rating scores and essay scores on the one hand and GPA on the other hand. One reason for this dearth may be the fact that such tests are almost always specific to the institution and the results from them may not be considered sufficiently generalizable to justify publication.

One interesting development which may inject new interest into this area is the introduction by ETS in 1986 of the TOEFL Test of Written English, which is standardized, intended for non-native speakers and widely available. Most of the current attention of the ESL/EFL profession is focussed on concerns about its content validity (e.g. Raimes 1990) and this survey has yet to unearth any predictor studies which incorporate its scores.

10-2 Use of admissions screening ratings as predictors of GPA: TOEFL

In line with its definition of itself as a "general" English proficiency test for non-native speakers, ETS warns on page 19 of its 1992 "TOEFL Test and Score Manual" :

> "Do not use TOEFL test scores for predicting academic performance .... Numerous predictive studies, using grade point averages as criteria, have been conducted in the past. These studies have shown that correlations between TOEFL test scores and grade-point averages are often too low to be of any practical significance."

The facts certainly support this unequivocal statement. The best efforts of correlation studies researchers in the 1970's and later produced only disappointing evidence of TOEFL's value as a perfect predictor of future academic performance: observed correlations were generally well below $r = .35$ (See Hale et al. 1984, for a summary of these studies). Thus, in recent years, ESL/EFL researchers have started to heed the warnings from ETS and others and have focussed their attention instead on TOEFL's value as a predictor of academic *failure* by paying special attention to scores at the lower end of the observed scale for admitted students. A major objective of this effort has been the attempt to determine within and across specific institutions TOEFL score levels which correlate highly with poor or mediocre academic performance. For the sake of convenience these hypothetical score levels will be referred to as "threshold levels".

The effort to find threshold levels within individual schools and institutions, let alone across institutions, has met with only very limited success because of factors that have been already mentioned: student populations at virtually all higher educational institutions which require the submission of TOEFL scores are pre-selected; there are very significant uncontrolled, and some uncontrollable, variables which mediate the relationship between English proficiency and academic performance. Among graduate institutions in the U.S.A. the pre-selection of the international student population is particularly severe. An ETS survey of over 300 U.S. graduate institutions (ETS 1992a) showed that 85% of them regarded a TOEFL score of 550 as a minimum admissions cut-off and that the 10% who employed a lower cut-off of between 500 and 550 coupled it with restrictions on course load or with enrollment in an English language program or with both. Only about 5% considered conditional admission to students with TOEFL scores lower than 500.

One way to grasp the significance of the 550 cut-off score is the fact that ETS calculates it represents the sixtieth percentile for all examinees who indicated that they were applying for admission to graduate schools (ETS 1992a). Given the combination of the pre-selection and various uncontrolled variables (see below), it hardly comes as a surprise to find that most studies of international student performance at U.S. graduate institutions concluded that the correlations with TOEFL scores were so low that the TOEFL cut-off for admissions purposes was safely set above any hypothetical "threshold level".

In many of these studies of international graduate student performance the correlations are indeed so low, between r = zero and r = .2 (e.g. Hale et al. 1984) that the cynical observer could be excused for wondering if the safe threshold level for most graduate institutions is not in the 500-600 TOEFL range at all, but much lower down the scale. Before jumping to such a conclusion, however, the possibility that these studies have not made allowance for mediating variables will have to be taken into account. Furthermore, if it can be shown that these variables have tended to bias the correlational norm which can be derived from these studies in one direction or the other, an operational distinction between "observed" and "true" correlations will have to be formulated and applied to all attempts to compare the results from any one institution with the norm. For this reason a critical analysis of at least a sample of the research into TOEFL as an academic performance predictor is necessary.

As Perry (1988) points out, there have been a lot of studies into the relationship between the first semester or first year GPA of international graduate students at U.S. colleges and universities and English proficiency as measured by the TOEFL Test, but well into the 1980's most of these were plagued by very small subject populations and other design problems. This is a great pity because the old 5-part TOEFL in use up until 1979 included separate sections on "vocabulary" and "writing ability" in addition to "listening comprehension", "structure" and "reading" and, thus, provided a much broader assessment of English proficiency than the current 3-part TOEFL. Moreover, it has to be pointed out that, although the authors of more recent individual studies seem to have learned from their predecessors' mistakes, most of the surveys which attempt to summarize results from different institutions still continue to ignore important variables within and across institutions that need to be taken into account before making comparisons.

Typical of recent surveys of studies into the relation between TOEFL scores and first semester or first year GPA is the one published by Graham in 1988. This survey summarizes the conclusions of a number of reports devoted to first semester or first year graduate GPA across a range of disciplines and indicates a range of correlations between r = almost zero and r = .40 (r^2 = almost zero to .16), though the mean is closer to the lower end of the range. Although this survey contains no summary of the descriptive statistics, it can be ascertained from an earlier survey report (Hale et al. 1984) that the TOEFL scores for the students in the studies reported by Graham range from as low as 400 to 600+, though 480 seems to represent the typical base score. In this survey not even the possible existence of mediating variables is recognized. It appears from the summaries by Hale et al. (1984) of the same studies that the individual studies themselves failed to make any allowance for intervening variables. Among the most significant of these variables are the internal ones of course load per semester and faculty differences in grade distribution and the largely cross-institutional ones of type of major and the language and cultural background of the students.

Most of these variables are highly familiar to most educators and need no further explanation, but the variability in course load per semester may be less familiar. Graham omits any reference to credit load per semester as an important variable in the performance of international students. There are indications in the survey by Hale et al. (1984) that there is such variability in U.S. graduate institutions. On top of this, in an excellent study by Light, Xu and Mossop (1987) which is referred to in detail below, there is clear evidence that the variability correlates significantly with the TOEFL measure of English proficiency. This leads to the suspicion that where international students with relatively weak English proficiencies enjoy some control of their semester course load they will sometimes reduce it in order to compensate for some of their linguistic disadvantages. It is important to bear this variability in mind when making comparisons with the MBA program at IUJ which imposes an extremely heavy and inflexible first year credit load on all students, regardless of their linguistic status.

One technical variable to bear in mind is the choice of first semester or first year grades as the basis for the calculation of GPA. Most of the studies cited by Graham and most of the studies cited in the much larger summary by Hale et al. are based on first semester GPA. Perry (1988:47), on the other hand, argues that correlational studies based only on first semester GPA "may not be tapping the potential source of variance available in cumulative GPA".

Of the individual studies on the academic performance of graduate international students produced in recent years, two of the most useful are those written by Perry (1988) and Light, Xu, and Mossop (1987).

Perry's study investigates performance at two U.S. universities, of which the largest study is the one at the University of Wisconsin with a subject population of 393 international graduate students enrolled in various arts and sciences programs. The descriptive statistics show that the mean TOEFL score for international students at the University of Wisconsin was 563.2, with a score range between 475 and 677. GPA was calculated over the first year or the full course. The correlation between TOEFL and GPA was weakly significant at $r = .18$ and $r^2 = .032$ ($p < .05$). Unfortunately at press time the portion of Perry's study which details grouping according to major, native language, and the like is unavailable. This information will be considered in a subsequent revision. In the meantime, there are two interesting results which are reported in the incomplete version of Perry's study. The first is the finding that that the gross correlations between TOEFL scores and academic failure across the total student population are somewhat unreliable because the results from the University of Wisconsin (UW) study ($N = 393$) contradict the results from the other, University of Minnesota (UM), study. The former study showed that in the 475-499 TOEFL band the mean GPA was the lowest for all bands at 3.36, whereas the latter study showed that in the same 475-499 band the mean GPA of 3.55 was actually higher than the mean GPA of the two TOEFL bands above it. Similarly inconclusive results are reported for the very high TOEFL bands above 600, though their overall mean of about 3.6 is slightly above the mean of around 3.5 for the TOEFL bands between 475 and 599. The second finding is that the Reading and Vocabulary section of the TOEFL correlated most highly with GPA at both schools ($r = .23$ at UW and $r = .43$ at UM), whereas the Listening Comprehension Section correlated the least ($r = .11$ at UW and at $r = .2$ at UM), and that in a regression analysis using the section scores, the score for listening comprehension added nothing to the observed relationship (Perry 1988). Because of

the large subject populations of Perry's two studies, this observation provides reliable confirmation of similar ones by a number of other researchers including Gue and Holdaway (1973) and Shay (1975).

Light, Xu, and Mossop's study was of 376 international graduate students at the State University of New York at Albany (SUNYA). It showed a weak (r = .14, r^2 =. 0196), though statistically significant, correlation between TOEFL and first semester GPA from a variety of arts and sciences courses. Despite the absence of an analysis of relationships with the TOEFL part scores, this study has four features which make it especially valuable.

The first feature is the wide spread of English proficiencies among the graduate students. Although the mean TOEFL score was 561, the range was a broad one from 400 to 677 in which 283 students had TOEFL scores of 530 or above and 93 students had scores below 530. Such a wide spread is unusual for a U.S. university these days when admissions policies mean that the typical range for international graduate students starts at 500 and above.

The second feature is a clear recognition of course load as a variable which might be related to English proficiency. As Tables 2 & 3 below show, TOEFL correlated more significantly with credit hours than with GPA.

Table 2

Mean number of credit hours earned according to TOEFL score range at SUNYA (F = 3.76, p <.005) (From Light, Xu & Mossop:1987)

TOEFL Score Range	*N*	Credit hours earned	
		M	*SD*
400-529	93	9.38	3.22
530-549	51	9.45	4.13
550-569	72	9.95	3.46
570-599	81	10.74	2.72
600-680	79	10.90	2.95

Table 3
Mean first semester GPA according to TOEFL score range at SUNYA ($F = 3.22, p < .01$) (From Light, Xu & Mossop:1987)

TOEFL Score Range	*N*	GPA Mean*
400-529	93	3.38
530-549	51	3.39
550-569	72	3.24
570-599	81	3.41
600-680	79	3.55

* overall GPA mean = 3.40 (*SD* = .54) GPA range = 2.25-4.00

Overall, for the TOEFL scores the SUNYA study found a correlation with GPA of $r = .14$ ($p < .05$) and with credit hours earned of $r = .16$ ($p < .01$).

The third valuable feature is that the study shows that the linguistic heterogeneity of the subject population- a heterogeneity which is typical of English medium graduate schools- represented a very significant uncontrolled variable. "European" students, for example, who had a mean TOEFL score just above the mean TOEFL score of "Korean/Japanese" students (569 vs. 561), nevertheless achieved a mean GPA well above that of the latter group (3.51 vs. 3.27).

The fourth feature is that observed correlations with GPA were higher for students enrolled in humanities, fine arts and social sciences programs ($r = .21$) and lower for those enrolled in science, mathematics and business programs. Interestingly, 43 of the SUNYA students were entered in the MBA program. For them the mean TOEFL score was 586 and the observed correlation between TOEFL and first semester GPA was $r = .02$ ($p < .05$). One probable explanation for this low correlation is the observation that the international students had a very low mean GPA of 3.13 which "is considerably lower than most other majors and lower than the overall mean GPA [of 3.40]". It is not clear whether the business students had any control over their course load.

The authors of this study concluded that the admissions procedure at the institution was restricting admission of students with TOEFL scores below 550 to those "with unusually promising academic ability". Such a restriction will necessarily depress observed correlations between English proficiency and academic performance. It is also clear that the other three variables identified in the SUNYA study tend to depress the same observed correlation. Though it will never be possible to determine with precision what the true correlation at SUNYA should be, it is certain that it must be higher than the observed one.

The only study that has ever convincingly showed moderately strong correlations between TOEFL and GPA was conducted by Gue and Holdaway in 1973. It is distinguished by three unusual characteristics which would tend to bias the correlation towards a true one. These are a low mean TOEFL score, uniformity of major and uniformity of language background. The subject population comprised 123 Thais who were admitted by the Education Department of the University of Alberta during the period 1967-1970 and had a mean TOEFL score of 424.6 at the start of the pre-matriculation summer intensive English programs. At this low

proficiency level their TOEFL scores and the GPA of their one year graduate education course correlated at $r = .49$, with the part scores for "Structure" and "Writing Ability" (in the old 5-part TOEFL) correlating the most at $r = .51$ ($p < .01$).

Within the parameters of this critical survey, some tentative conclusions can be drawn about the true norms of correlations between English proficiency, as measured by the TOEFL test, and graduate academic performance. The first is that the correlations found by the Gue and Holdaway study represent a strength of relationship which is below the "safe" threshold level. The reported mean TOEFL score of 424.6 is somewhere between the fifth and ninth percentile for all 561,629 TOEFL examinees who between 1989 and 1991 indicated they were applying to graduate programs in North America (ETS 1992a). At such a low mean proficiency level it is very likely that lack of proficiency was handicapping the academic performance of the students. The second conclusion is that the observed correlational norm among most studies of r = zero to $r = .2$ is most likely below the true correlation. Of course, the really significant correlational value is the r^2 value which expresses the amount of shared variance between two variables and is used as the measure of perfect prediction. It is a peculiar property of a such a squared ratio that it remains relatively small when the raw correlation is below $r = .4$, but suddenly shoots up above a raw correlation of about $r = .45$. Since this value is also close to the value reported in the Gue and Holdaway study, it follows that the appearance in the IUJ study of any raw correlation between TOEFL and GPA which is close to this value would be a cause for concern.

10-3 Use of admissions screening ratings as predictors of GPA: GMAT

Up to this point the literature search has not revealed any studies that are specifically devoted to investigating the relation between the GMAT scores and GPA of international students enrolled in MBA programs in the U.S.A. The data which is available refers to the relationship for all enrolled students. Through its offshoot, the Graduate Management Admissions Council (GMAC), ETS claims that 38 validity studies conducted in the period 1990-1991 show the following correlations (GMAC 1991):

Table 4

The relationship between first year GPA in 38 MBA programs and GMAT scores and undergraduate GPA for all students (From GMAC 1991)

First Year GPA = Dependent Variable

Independent Variables	*r*	*Median r*
GMAT-V+Q	ranges from .16 to .62	.35
GMAT-V+Q+UGPA	ranges from .21 to 67	.43

10-4 Use of admissions screening ratings as predictors of GPA: Undergraduate record

Little research into the relationship between the undergraduate record and graduate performance of international students has been published. One good reason for this is the difficulty in generalizing within institutions, let alone across institutions, because of the enormous problems in equating grades and reports from different educational cultures and types of institution. Another reason is probably the depressive effect on observed correlations of subject pre-selection and broad-band calibration. For native speakers in the U.S.A. Perry (1988) cites studies which show simple correlations ranging from .13 to .37 and later indicates that at the Universities of Wisconsin and Minnesota the observed correlations for international students have been below .20.

Japanese undergraduate records are not subject to cross-cultural variation, but the relatively relaxed attitude to undergraduate education in Japan means that the value of undergraduate records as a graduate performance predictor might not be very high.

11 English proficiency scores as GMAT moderators: Empirical findings in the literature

The research into the GMAT moderator function of English proficiency scores is limited to the standardized test, TOEFL. A definitive study by Powers in 1980 of 5,781 non-native speakers who took both tests within the two year period, 1977-1979, showed the following correlations:

Table 5

Simple Correlations between GMAT and TOEFL Scores (From Powers 1980:7)

GMAT Scores	Simple Correlation with TOEFL
GMAT-Verbal	.71
GMAT-Quantitative	.39
GMAT-Total	.66

Table 6 shows the correspondence between TOEFL scores and predicted GMAT scores at various TOEFL levels, when computed by regression analysis, using a linear fit.

Table 6
GMAT Scores Predicted at Various TOEFL Levels (From Powers:1980:10)

TOEFL Score	GMAT-V[a]	GMAT-Q[b]	GMAT-Total
450	7.5	22.8	290.4
500	12.0	25.5	341.5
550	16.5	28.3	392.5
600	21.0	31.0	443.6
650	25.5	33.8	494.7

a.*Standard Error of Measurement* (one-directional) for GMAT-V = 5.9.
b.*Standard Error of Measurement* for GMAT-Q not given (but must be very high)

12. Intercorrelations among English proficiency measures: empirical findings in the literature

The best evidence identified so far is from Pike (1979) who looked for correlations between the parts of the old five-part TOEFL and experimental interview and essay subtests. The essay subtest required a more elaborate written product than the "Writing Ability" section of the old 5-part TOEFL. It should be noted that the measurement scales were finely calibrated. The results summarized here relate to the 192 Japanese subjects who participated in the project. Intercorrelations with the five parts ranged between .59 and .82 for interview and between .72 and .81 for the "form" rating of the essays. Essay "content" correlated between .45 and .64. It is, however, unwise to draw sweeping conclusions from these observed correlations since interview and essay tests for admissions screening purposes are specifically designed to serve the needs of a single institution or school.

13. Methodology

Graduate transcripts of 61 Japanese students and 38 other students enrolled in the School of International Management at IUJ for the period of Fall 1989 to Spring 1992 were provided by the GSIM Office. After excluding all grades for language courses, GPA was calculated by hand for all first year content courses. The GPA range extends from 1 to 4. The advantage of focussing on first year GPA is that, unlike the second year courses, the first year courses are compulsory for all students and provide grading data which is less subject to uncontrolled variables than the second year data. It is often argued that first semester grades only should be included in the GPA calculation on the grounds that grades from following semesters can confound "English language proficiency ....prior to beginning graduate study with English proficiency gained during study...." (Light, Xu and Mossop 1987:54). Others, such as Perry, have argued first semester grades are subject to distortion because of psychological factors and that a GPA data base gathered over a longer time span is more likely to reflect the true abilities of the student. Both these arguments have merits, but the fact that nearly all the Japanese students attend a ten week intensive English program prior to matriculation weakens the force of the first argument in relation to this study. Furthermore, GPA data collected over a longer period is more

likely to be helpful to professors and administrators in the IM School who tend to be interested in the performance over the long term of admitted students.

An attempt has been made to determine what kind of quantitative bias exists in the first year MBA courses at IUJ. The conclusion that the first year IM courses have a strong quantitative bias was reached after collating the results of a questionnaire that was returned in time for this paper by four full-time professors in GSIM.

All the score data for the admissions screening measures was obtained directly from the IM Office. This data also includes the scores for candidates admitted in the Fall of 1992. Although the scoring criteria for the in-house screening measures have not changed very much during the period under investigation, there have been some changes in the way scores are calibrated, making it necessary to convert some scores to fit the five band calibrations which have become the norm since 1992. The conversions that have been performed affect the percentage scores allocated to the essay and interview tests in 1989 and the three-band measures used to rate undergraduate achievement in 1990 and 1991. The GSIM uses GMAT percentile information in making admissions decisions. In order to permit correlation analyses between scores of a uniform type, this study uses GMAT interval scores.

The "Statview SE+ Graphics" [TM] software package was used to analyze the data for this study. The total number of Japanese subjects entered in the data base is 98, of which 9 are dummy subjects, leaving a total of 89 subjects which can be used for descriptive statistics involving admissions scores only. Of this figure of 89, 28 represent students admitted in 1992 who have yet to complete the first year of studies. The remaining 61 students include 60 who obtained credits for all the first year courses during the period 1989-1992. Thus, for analyses involving GPA the maximum number of subjects is 60.

The descriptive statistics for the standardized tests, TOEFL and GMAT, and for GPA are presented in Table 7 below. Descriptive statistics for the other screening measures in use during the period 1989-1991 are included in the Appendix. It should be noted that the mean TOEFL score of the group of 556.7 is above the mean of 529 reported by the Educational Testing Service (1992) for all TOEFL examinees who indicated they were applying for admission to graduate level courses. The mean GMAT scores for all GMAT examinees reported by the Graduate Management Admissions Council (1991:12-13) for the period 1988-1991 are GMAT Total: 494, GMAT Verbal: 27, GMAT Quantitative: 30.

Table 7
Mean TOEFL, GMAT and first year GPA data for students enrolled in GSIM, 1989-1991

	N	%	TOEFL	GMAT	GMAT-V	GMAT-Q	GPA
Japanese	*60*	*60.6*	556.7	486.9	17.8	39.7	2.64
(SD)			*(35.6)*	*(61.6)*	*(5.0)*	*(5.1)*	*(0.49)*
All*	*99*	*100.0*	568.6	525.1	22.9	39.8	2.74
(SD)	*86 for TOEFL*		*(39.8)*	*(81.1)*	*(9.1)*	*(5.5)*	*(0.46)*

*(60 Japanese, 26 Other Non-Native Speakers, 13 Native Speakers)
Note: for Japanese subjects the TOEFL score range = 487-633 and the GPA range = 1.64-3.86
for all subjects the TOEFL score range = 487-650 and the GPA range = 1.64-3.86

14. Hypotheses

The hypotheses which inform the data collection and organization of this IUJ study fall into three areas: hypotheses about predictors of GPA, hypotheses about the role of English proficiency scores as moderators of GMAT, and hypotheses about redundancy among all the admissions screening measures.

14-1 Hypotheses about predictors of GPA

(1) TOEFL will correlate moderately at a level somewhere above the norm observed in the roughly comparable studies cited above because of the relative homogeneity of the subject population, this population's numerical predominance in GSIM and because GPA is a controlled variable. It will correlate well below the level reported by the Gue and Holdaway study because of the much higher mean TOEFL score reported for this study.

(2) Essay and Interview Tests will not correlate at all in a significant way because of the pre-selected population and the rough calibration of the measuring devices.

(3) GMAT will correlate fairly strongly because it is specifically designed to serve as a predictor even for pre-selected populations. Given the strong bias towards quantitative courses in the first year, the Quantitative section of the GMAT should correlate even more strongly.

(4) The other screening scores will correlate weakly or not at all because of the pre-selected population and the rough calibration of the measuring devices.

14-2 Hypotheses about English proficiency scores as moderators of GMAT

(1) TOEFL will moderate GMAT-V to a significant degree

(2) TOEFL will moderate GMAT-Q less significantly

14-3 Hypotheses about redundancy among all the admissions screening measures

There will no observed overlap of any great significance (i.e. above $r^2 = .5$) because of the broadly calibrated intervals of the majority of the measuring devices. Powers' 1984 study of the relationship between TOEFL and GMAT leads to a similar hypothesis that there is no significant redundancy among these standardized measures.

15. Results

The results are presented in the same order as the hypotheses in Section 14 above.

15-1 Predictors of GPA

A simple correlational analysis of the relation between GPA as the dependent variable and the six admissions screening measures in use over the whole period 1989-1991 was first performed. The relationship between GPA and the other four measures in use only from 1990 was analyzed separately. The results from these two analyses are presented in Tables 8 & 9.

Table 8

Simple correlations between GPA as the dependent variable and six GSIM independent admissions variables for the period 1989-1991. Japanese students only. $N = 60$

Screening Measure	r	r^2
GMAT-Q	.634*	.402
GMAT Total	.566*	.321
GMAT-V	.380**	.144
TOEFL	.286***	.082
Interview - English Proficiency	.237	.056
Essay - English Proficiency	.111	.012

* significant at $p < .0001$ ** significant at $p < .005$ *** significant at $p < .025$
other correlations insignificant at $p < .05$

Table 9

Simple correlations between GPA as the dependent variable and four GSIM independent admissions variables for the period 1990-1991. Japanese students only. N = 33-42

Screening Measure	r	r^2
Essay - Contents	.313*	.098
Undergraduate Transcript	.236	.056
Undergraduate Institution	.130	.017
Interview - General	.087	.007

* significant at $p < .05$ all other correlations insignificant at $p < .05$

The r^2 figure is the most pertinent as it describes the shared variance of the dependent and independent variable.

Since full regression analyses with a subject population of less than 100 are unreliable (personal communication from Perry), regression analysis is limited (a) to identifying the statistical significance of the shared variance between the dependent variable, first year GPA, and all the independent variables lumped together as a single block, (b) to identifying the degree of shared variance which is not accounted for by the main contributor, the GMAT Quantitative scores, and (c) to identifying the shared variance which can be accounted for by the three English proficiency measures. Regression analysis involving the independent variables which cover the period 1990-1991 could not be performed because of the small N of 33-42. GMAT Total scores are omitted as they inter-correlate perfectly with the combination of GMAT-V and GMAT-Q. Results from the limited regression analyses that were performed are summarized in Tables 10 & 11 below.

Table 10

Multiple regression summary table: All GSIM admissions variables 1989-1991. Japanese students only.

Dependent Variable = GPA $N = 60$

Order	Independent Variable	Standard Beta	Multiple r	r^2	p
1	GMAT-Q	.608	.634	.402	.0001
2	Four Others	-.08 to .129	.659	.434	insignificant (.57 to .23)

Table 11

Multiple regression summary table: Three GSIM English proficiency admissions variables 1989-1991. Japanese students only.

Dependent Variable = GPA $N = 60$

Independent Variables	Standard Beta	Multiple r	r^2	p
TOEFL	.222	.319	.102	.11
Essay (English Proficiency),	.038			
Interview (English Proficiency)	.143			

The high value for p in the above table serves as a warning that it would be best to accumulate a larger N before conducting any further regression analyses.

In view of the relatively weak correlations between the English proficiency measures and first year GPA, Table 12 shows a comparison of means within score bands which is restricted to TOEFL, the proficiency measure with the highest correlation.

Table 12
Mean GPA earned by Japanese students according to TOEFL score range, 1989-1991

TOEFL Score Range	*N*	%	TOEFL Mean + *(SD)*	GPA Mean + *(SD)*
480-540	*19*	*31.7*	518.7 *(16.4)*	2.57 *(0.47)*
541-573	*24*	*40.0*	555.9 *(9.2)*	2.54 *(0.47)*
574-640	*17*	*28.3*	602.3 *(16.8)*	2.86 *(0.49)*

15-2 English proficiency measures as moderators of GMAT

A similar analysis to Powers' was conducted to see if TOEFL levels moderate GMAT scores in the same way that Powers finds (See Table 5 above). For this analysis the scores from 1992 are also included. Table 13 presents these results.

Table 13
Mean GMAT scores earned by Japanese students according to TOEFL score range, 1989-1992

TOEFL Score Range	*N*	%	TOEFL Mean + *(SD)*	GMAT-V Mean + *(SD)*	GMAT-Q Mean + *(SD)*
480-510	*7*	*7.1*	500.0 *(10.1)*	13.5 *(3.8)*	35.3 *(4.5)*
511-540	*23*	*23.2*	528.2 *(7.8)*	15.1 *(3.8)*	38.7 *(5.6)*
541-573	*35*	*35.4*	557.7 *(8.8)*	17.6 *(3.6)*	41.0 *(4.4)*
574-640	*24*	*24.3*	599.2 *(15.6)*	20.6 *(5.8)*	41.6 *(5.4)*

In view of the low intercorrelations between GMAT scores and other English proficiency measures (See Tables 14 & 15 below), no other analysis of GMAT score moderation was attempted.

15-3 Evidence of redundancy among all the admissions screening measures

Tables 14 & 15 present the results from a simple correlational analysis of all the screening measures. These show little observible evidence of significant overlap among the measures. Once again, however, it must be stressed that, because of the restricted distributions of scores within the measures and/or their broad calibrations, the correlations between many of the measures may well be understated.

Table 14

Intercorrelations between six GSIM admissions measures in use 1989-1992 (N =89)

Correlation Matrix for Variables: $X_1 \ldots X_6$

	Toefl	GMAT	Verb	Quan	Ess E	I E
Toefl	1					
GMAT	.473	1				
Verb	.486	.844	1			
Quan	.333	.844	.441	1		
Ess E	.124	-.01	.086	-.113	1	
I E	.369	.131	.124	.111	.045	1

Table 15

Intercorrelations between ten admissions measures in use 1990-1992 (N = 52)

Correlation Matrix for Variables: $X_1 \ldots X_{10}$

	Toefl	GMAT	Verb	Quan	Ess E	I E	Ess C	I G	
Toefl	1								
GMAT	.548	1							
Verb	.517	.859	1						
Quan	.414	.832	.442	1					
Ess E	.059	.079	.135	-.037	1				
I E	.327	.084	.093	.064	.103	1			
Ess C	.241	.046	.007	.082	.023	.286	1		
I G	.156	.017	-.115	.166	-.072	.637	.371	1	U I
U I	.227	.203	.133	.217	.033	-.135	-.035	.132	1
U T	.23	.122	.168	.028	.016	.117	.283	.007	-.109

The only overlap of any significance is between the Interview rating for English proficiency and the General Interview rating.

16. Interim conclusions

16-1 Factors which constrain the interpretation of the scores from the English proficiency components of the admissions battery

The statistical characteristics of the English proficiency measures used in the admissions screening procedure mean that they are subject to a very restricted set of interpretations. The English Essay and Interview measures are so broadly calibrated that confident interpretations can yield only broad descriptors such as "weak" and "strong". Furthermore, the high degree of pre-selection in the subject population means that all the English proficiency measures are measuring but a short part of the proficiency spectrum and are thus susceptible to distortion by error and imprecision. For this reason alone, it would be wise to regard them as simple threshold measures which, in combination with other indicators, suggest whether or not a candidate has the potential to benefit from the program of study.

Conclusions 16-2, 16-3 and 13-4 are largely based on empirical investigations which are still in progress. They should not be regarded as final conclusions.

16-2 Admissions screening English proficiency scores of Japanese students in GSIM as academic success predictors and as academic aptitude moderators

The statistical characteristics of the English Essay and Interview measures make it inherently unlikely that they would predict academic success or moderate standardized academic aptitude scores and this propensity is reflected in the results. The results also bear out the Educational Testing Service's warning that the TOEFL Test should not be regarded as a measure of English academic aptitude which can help to predict academic success. Indeed, all the indicators from the literature lead to the conclusion that it is precisely when TOEFL scores do correlate fairly significantly with academic success at more than about $r = .40$ that there are grounds for concern that the English proficiency threshold has been set at too low a level. As regards TOEFL's possible role as an English academic aptitude moderator, the results correspond fairly well with those reported in Powers' definitive study. There is a distinct possibility that in cases where an applicant's TOEFL score is relatively low the GMAT score will be depressed significantly below the GMAT score the same applicant would achieve with a higher TOEFL-measured English proficiency. This observation applies quite strongly in respect of the GMAT Verbal score level, but more weakly in respect of the GMAT Quantitative score level.

16-3 Admissions screening English proficiency scores of Japanese students in GSIM as predictors of academic success or failure

The English proficiency data that are available cannot be used to provide consistent predictions of academic failure and, thus, there is no evidence yet that lack of English proficiency is unduly handicapping the academic performance of the Japanese students. Japanese students in the lowest 480-540 TOEFL band are able to obtain academic grades which are just as high as those of Japanese students in the 541-573 TOEFL band. There is some evidence, however, that relatively high TOEFL scores correlate with academic success. Japanese students with English proficiencies in the highest 574-640 TOEFL band obtained a significantly higher GPA.

The English proficiency data can, at best, account for only a minor share of observed academic performance. By far the most significant source of shared

variance with first year GPA is the GMAT Quantitative score, which is moderated only modestly by TOEFL-measured English proficiency. Thus far, the three English proficiency measures used for admissions screening have a shared variance with first year GPA of no more than 10%. This is at the high end of the observed range that it is typical for international students at graduate schools in the U.S.A.. When the factors that have tended to depress the observed correlations in the U.S. studies, such as subject population and course heterogeneity and the probable lack of uniformity in grade distribution, are taken into account, this shared variance would appear to be very much within the true range. Furthermore, even this modest shared variance almost disappears when superimposed on the overlap between first year GPA and GMAT-Q.

For now, at least, the evidence that is available suggests that, in combination with the other screening measures and the pre-matriculation Intensive English Program, the English proficiency thresholds used for admissions screening are not so low that they prevent a large number of the Japanese students from realizing their academic potential in the MBA program at IUJ. Significant changes in GSIM, such as changes in the composition of the student enrollment or in the content of the curriculum, however, would necessitate a completely new study.

16-4 Evidence of redundancy among the English proficiency scores and within the entire admissions screening battery

There is no evidence of any significant redundancy among the English proficiency scores. It has to be remembered, however, that the same statistical characteristics that make it unlikely that the English Essay and Interview scores will correlate with academic success also make it unlikely that they will correlate with TOEFL and each other. Pike's 1979 study, using much more finely calibrated measuring scales, did reveal fairly high correlations between TOEFL and experimental interview and essay tests, but it is impossible to equate these experimental tests with their IUJ counterparts since both sets of tests are unstandardized. In this regard it is interesting to note that the correlation between TOEFL scores and the highly calibrated reading and writing scores in the tests given to the Japanese MBA students upon entry to the Intensive English Program at IUJ show a mean correlation of about $r = .25$, higher than the correlation for the admissions tests, but hardly indicative of a significant overlap.

The only hint of redundancy among the full battery of screening measures concerns the scores for the General and English proficiency evaluations of the Essay. Yet again, however, it should be stressed that the statistical characteristics of the screening measure scores may be having the effect of making the relationships seem weaker than they really are.

17. Future research

In addition to collecting more of the same data from a larger subject population a search through the non-linguistic literature sources would probably be useful. Future research, for the reasons explained in an earlier section of this paper, should also try to incorporate the part scores from the TOEFL Test. With part scores being available from 1991, it should be possible to accumulate sufficient data to perform separate part score analyses in 1994 or 1995. Another dimension which might be worth exploring is that of targeted academic English proficiency. Scores from the

Text Skills for International Management (TSIM) component of the Intensive English Program exist in a fairly standardized form from 1990, so that by 1994 or 1995 it should be possible to perform analyses which factor in evaluations of academic English skills and behaviors which are directly related to the demands of the MBA program, yet by design entirely avoid quantitative skills work. Provisional studies using a small subject population of under 50, for example, have indicated that the exit scores from TSIM correlate with first year GPA more significantly than the English proficiency admissions battery.

References

American Association of Collegiate Registrars and Admissions Officers. 1971. AACRAO-AID participant and placement study. (Report to the Agency for International Development, U.S. Department of State). Washington, DC: USAID. [Cited in Hale et al., and in Graham]

Brown, J.D. 1988. Understanding research in second language learning. Cambridge: Cambridge University Press. Chapter 10, 126-153.

Educational Testing Service. 1992a. TOEFL Test and Score Manual. Princeton, NJ: Educational Testing Service.

Educational Testing Service. 1992b. Bulletin of information for TOEFL/TWE and TSE. Princeton, NJ: Educational Testing Service.

Graduate Management Admission Council. 1991. Guide to the use of GMAT scores. Princeton, NJ: Educational Testing Service.

Graduate School of International Management. 1992. MBA Student Handbook 1992-1993. International University of Japan: Yamato-machi, Niigata-ken.

Graham, J.C. 1987. English language proficiency and the prediction of academic success. TESOL Quarterly 21: 505-521.

Gue, L.R., & E.A. Holdaway. 1973. English proficiency tests as predictors of success in graduate studies in education. Language Learning, 23: 89-103. [Cited in Hale et al.]

Hale, G.A., C.W. Stansfield, & R.P. Duran. 1984. Summaries of studies involving the Test of English as a Foreign Language, 1963-1982. (TOEFL Research Report 16). Princeton, NJ: Educational Testing Service.

Hatch, E., & H. Farhady. 1982. Research design and statistics for applied linguistics. Rowley, Mass.: Newbury House.

Light, R.L., M. Xu, & J. Mossop. 1987. English proficiency and academic performance of international students. TESOL Quarterly 21: 251-261.

Oller, J.W., & B. Spolsky. 1979. The Test of English as a Foreign Language. In B. Spolsky (ed.), Some major tests. Advances in language testing series: 1. Papers in applied linguistics, 92-100. Arlington, VA: Center for Applied Linguistics.

Perry, W.S. 1988. The relationship of the Test of English as a Foreign Language (TOEFL) and other critical variables to the academic performance of international graduate students. Unpublished Ph.D thesis, University of Minnesota.

Pike, L. 1979. An evaluation of alternative item formats for testing English as a foreign language. (TOEFL Research Report 2). Princeton, NJ: Educational Testing Service. [Cited in Hale et al.]

Powers, D.E. 1980. The relationship between scores on the Graduate Management Admission Test and the Test of English as a Foreign Language (TOEFL Research Report 5). Princeton, NJ: Educational Testing Service.

Raimes, A. 1990. The TOEFL Test of Written English: causes for concern. TESOL Quarterly 24: 427-442.

Shay, H.R. 1975. Affect [sic] of foreign students' language proficiency on academic performance. Dissertation Abstracts International 36, 1983A-1984A. (University Microfilms No. 75-21, 931) [Cited in Hale et al.]

Appendix

Descriptive statistics for six admissions screening measures, 1989-1991

(1) First year Grade Point Average

X_1 : GY

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
264.017	48.567	6.27	2358.762	18.395	60
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum of Sqr.:**	**# Missing:**
164	386	222	15841	4321455	7

Also: Median = 257, Skewness .201

X_1 : GY

Bar:	From: (≥)	To: (<)	Count:	Percent:	
1	164	185	3	5%	
2	185	206	3	5%	
3	206	227	8	13.333%	
4	227	248	7	11.667%	
5	248	269	13	21.667%	
6	269	290	4	6.667%	
7	290	311	14	23.333%	-Mode
8	311	332	3	5%	
9	332	353	2	3.333%	
10	353	374	2	3.333%	
11	374	395	1	1.667%	

(2) GMAT Total

X_1 : GMAT

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
486.885	61.605	7.888	3795.137	12.653	61
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum of Sqr.:**	**# Missing:**
390	680	290	29700	14688200	6

Also: Median = 480, Skewness = .569

X_1 : GMAT

Bar:	From: (≥)	To: (<)	Count:	Percent:
1	390	420	7	11.475%
2	420	450	10	16.393%
3	450	480	13	21.311%
4	480	510	7	11.475%
5	510	540	13	21.311%
6	540	570	4	6.557%
7	570	600	4	6.557%
8	600	630	2	3.279%
9	630	660	0	0%
10	660	690	1	1.639%

(3) GMAT-Q

X_1 : Quant

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
39.633	5.076	.655	25.762	12.806	60
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum of Sqr.:**	**# Missing:**
29	50	21	2378	95768	7

Also: Median = 39.5, Skewness = .185

X_1 : Quant

Bar:	From: (≥)	To: (<)	Count:	Percent:	
1	28	31	3	5%	
2	31	34	3	5%	
3	34	37	11	18.333%	
4	37	40	13	21.667%	- Mode
5	40	43	12	20%	
6	43	46	11	18.333%	
7	46	49	5	8.333%	
8	49	52	2	3.333%	
9	52	55	0	0%	
10	55	58	0	0%	

(4) GMAT-V

X_1 : Verb

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
17.833	5.029	.649	25.294	28.202	60
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum of Sqr.:**	**# Missing:**
8	35	27	1070	20574	7

Also: Median = 17.5, Skewness = .794

X_1 : Verb

Bar:	From: (≥)	To: (<)	Count:	Percent:	
1	8	11	3	5%	
2	11	14	7	11.667%	
3	14	17	17	28.333%	- Mode
4	17	20	12	20%	
5	20	23	12	20%	
6	23	26	5	8.333%	
7	26	29	2	3.333%	
8	29	32	1	1.667%	
9	32	35	0	0%	
10	35	38	1	1.667%	

(5) TOEFL

X_1 : Toefl

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
556.672	35.57	4.554	1265.257	6.39	61
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum of Sqr.:**	**# Missing:**
487	633	146	33957	18978831	6

Also: Median = 553, Skewness = .196

X_1 : Toefl

Bar:	From: (≥)	To: (<)	Count:	Percent:	
1	487	502	4	6.557%	
2	502	517	3	4.918%	
3	517	532	7	11.475%	
4	532	547	9	14.754%	
5	547	562	13	21.311%	- Mode
6	562	577	8	13.115%	
7	577	592	5	8.197%	
8	592	607	3	4.918%	
9	607	622	7	11.475%	
10	622	637	2	3.279%	

(6) English Essay
(original 1-5 band multiplied by 10)

X_1 : Ess Eng

Bar:	From: (≥)	To: (<)	Count:	Percent:	
1	1	11	0	0%	
2	11	21	5	8.197%	
3	21	31	22	36.066%	
4	31	41	33	54.098%	- Mode
5	41	51	1	1.639%	

Also: Median = 40, Skewness = -.488

(7) English Interview
(original 1-5 band multiplied by 10)

X_1 : Int E

Bar:	From: (≥)	To: (<)	Count:	Percent:	
1	1	11	1	1.639%	
2	11	21	6	9.836%	
3	21	31	22	36.066%	
4	31	41	27	44.262%	- Mode
5	41	51	5	8.197%	

Also: Median = 40, Skewness = -.44