Estudios Ingleses de la Universidad Complutense
1999, n.º 7: 89-107
ISSN: 1133-0392

Is the English test in the Spanish University Entrance Examination as discriminating as it should be? [1]

Honesto HERRERA SOLER
Universidad Complutense

ABSTRACT

It is taken for granted that the aim of the Spanish University Entrance Examination is to discriminate among students accurately and reliably. A pythagorean analysis on a sample of 450 English tests to enter Spanish Universities is carried out to check whether this target is met. The dispersion of the scores, the skewness and the poor discriminative power, due both to the content and design, in the objective items of the test lead us to question the validity of the present design.

1. INTRODUCTION

1.1. Categorisation

Should the English test (ET) [2] in the Spanish University Entrance Examinations be categorised as a placement test or as a proficiency test? At first sight, because of the circumstances under which the test is taken and its intended aim, it might be considered closely related to a placement test, which is a widely known test type to screen foreign students entering a British university. As the ET is also taken as a screening test to enter a Spanish university, there is a tendency to consider it in similar terms. If, however, the question is given further consideration and we attempt to come up with a more accurate and academic answer, the ET should be categorised as a proficiency test. Among testers, a placement test is categorised as a criterion-referenced test, i.e., the focus is on how the students achieve in relation to the

material [3], while a proficiency test is seen as a norm-referenced test, which is used primarily to spread students out into a normal distribution so that their performances may be compared in relation to each other (Brown 1988). The aim of a placement test is to allocate students to one or another level while in a proficiency test discrimination is what really matters. In the case of the ET, the target is to discriminate as reliably as possible. The University Examination Board looks for an accurate score which enables the academic authorities to rank students according to their proficiency and, which at the same time, allows the students to make their choice of Faculty courses according to the score obtained. That is why the ET must be categorised within the range of types of proficiency tests.

1.2. Background

Research both on placement tests (Wall, Clapham and Alderson, 1994), and on ET in the Spanish University Entrance Examinations is scant: just cut off points, pass or fail percentages and little more in the media. On the other hand, literature on proficiency tests is quite abundant although there is still scope for further research in this field. Agreement is sought on the concept, terms, components, features and techniques among applied linguists but the prototypical proficiency test has not been found yet. Chastain (1988:49) studies the issue of the concept "to date, the profession has no acceptable definition of proficiency" whereas Harlow and Caminero (1990) focus their attention on the main features of a proficiency test. They document that researchers use a variety of terms and diverse components when assessing proficiency and this diversity makes it difficult to reach agreement on what constitutes a prototypical proficiency test. McNamara (1990) and Kenyon and Stansfield (1992) are also concerned with the problem of the components. From their perspective these components are mainly based on the variety of existing models. Alderson takes up this issue of model diversity and argues (1991:8) that "the profusion of competing and contradictory models, often with very slim empirical foundation, inhibits the language tester or applied linguist from selecting the best model on which to base his / her language test". Chalhoub-Deville (1997) also examines the relationship between theoretical models and operational assessment frameworks. In spite of the diversity of terms, components and models, there exists a point of agreement among researchers: there is a tendency among proficiency testers to structure the test in relation to the theoretical model of language competence referred to.

1.3. Model

What model underlies the ET? For the English Language Institutes (ELI), the purpose of a test is to identify students' language proficiency level since a lack of language skills or ability to communicate might cause students problems in their academic work in the departments where they study. The ET is also concerned with the students' performance. However, its aim is not only to identify but also to assess their language proficiency level. English is taken as a subject with a specific weight for the purpose of ranking students rather than for any considerations as to its being a useful tool for further studies. Upon analysing the Spanish University Examination Boards' designs as regards the subject, it will be observed that reading and writing skills rather than the oral dimension of communicative competence are highlighted. Consequently, the ET is not so much concerned with the communicative competence models (Canale and Swain, 1980, Canale,1983) as with one of the components of the communicative language ability model (CLA) (Bachman, 1990a): language competence. Within language competence the main issue is organisational competence, which, in turn, includes grammatical and textual competence, further broken down into grammar, lexis, reading comprehension and composition so as to provide a more detailed description of the construct. This is the operational framework most of the Spanish University Examination Boards use, though the issue regarding the nature of the components / items and the testing techniques may vary to some extent.

1.4. Features

This particular proficiency test, which the ET is —a norm-referenced test focused on the dispersion of scores— demands the same features as any other proficiency test:

1. Face validity: students' perception of whether the test is appropriate or inappropriate.
2. Content validity: whether tutors think that the programme content, in this study the COU [4] level, is represented in the test or not.
3. Construct validity: whether it really measures the ability / skill in reading and writing which the University Examination Board wants to measure.
4. Concurrent validity: the degree of correlation with the assessment of the tutors in the private or public educational institutions.
5. Reliability: the extent to which the results can be considered consistent and stable.

1.5. Scope

It is not within the scope of this study to go through each of these features exhaustively, but rather to focus specifically on those which may affect the ET design. It seems, among institutions, students, and evaluators, that much of the debate on the ET should be directed to the open items (OI) and especially to the essay, since they are considered subjective items. Arguments for and against can be expected on whether the number or the type of errors should be the main reference, or whether a more personal and original essay with some blatant mistakes should be rated worse / better than a conventional, simple and poor but standardised essay. Contrarily, it has always been assumed that there is no room for any sort of debate on those items which everybody labels as objective. It has been taken for granted that when assessing lexical or grammatical items through objective techniques there are no doubts, no dependency on the evaluator's mood and no problems with reliability among raters. Nevertheless, a follow-up of the so-called objective items appearing on the ET over the last few years shows that things are not as straightforward as they might seem to be. The dispersion of the scores in these items is quite distant from a normal distribution. The skewness here involved is not so much a problem of the raters as a problem of the content and design; in fact, scores are better spread out in the subjective than in the objective items.

1.6. Interest

The main concern of this research is the dispersion of scores in some items of the ET which may affect the content validity, or, at least, the validity in some cases on the grounds that they are not as discriminating as they should be. What we intend to carry out in this paper is a numerical approach to the scoring of the items rather than a technical approach to the structure and nature of the items themselves. Nevertheless, in this discussion comments on the design drawn from the data obtained cannot be ignored. The data analysed are the scores on the different items of the test, that is, the quantitative assessment of the proficiency in English shown by the students. Therefore, this study consists of a reading of the skewness, mean, analysis of variances and other statistics to see if this test accomplishes its goals: to spread scores out into a normal distribution (Tables III and V.c). Neither the topic of the reading comprehension, nor the components of the test are an immediate target in this analysis, though in the light of the data obtained some sort of revision for the content, level and design of the ST should be undertaken.

1.7. Hypothesis

An ET is expected to rank students' performance according to their level of learning and spread out students scores into a normal distribution. With the 1998 ET data provided, doubts about the accomplishment of its purpose must arise. Thus, the intention of this study is to test the hypothesis that the skewness of objective items could have affected the validity of this test.

2. METHOD

2.1. Conditions

The constraints of the ET performance of June 1998 are within the range of normality for this sort of tests: control of students, identification, administration of the test, invigilation, etc., are the same as hold for other subjects like Mathematics, Philosophy, History or Spanish Language. The time allowed is the only difference, since the modern language paper performance cannot exceed an hour. The evaluators, males or females, working either at the University or in Secondary Education, are asked to make the effort to score all examinations within the first five days after the date of the ET. Anonymity is maintained throughout the marking process. Raters do not know which educational institution students come from, nor the students' names. It is only once they have handed over the marked tests that they are allowed to know the educational institution they have assessed.

2.2. Subjects

For this study the scores given by 8 raters to the first 30 students they have marked are taken for analysis [5]. This number of students is considered a suitable figure for any statistical test and also ensures the appropriate representativeness for each educational institution in spite of its size. As all but one of the raters have assessed more than one private or public educational institutions, two marking lists [6] are taken from each of them and only one from the rater who assessed only one of these educational institutions. That means that the sample is taken from 15 different educational centres, with a total of 450 subjects evaluated. Randomness is taken for granted because each evaluator was given no more than 200 tests on no other criteria from the University Examination Board than that of distributing a similar number of examinations to each evaluator. Furthermore, the Administration did not know which raters would provide the data for this study; consequently the chances of being selected were the same for each test.

2.3. Components of the ET in the Spanish university entrance examination

The ET consists of five different items: a reading passage with two comprehension questions (open item, OI), two True / False questions (T/F), also based on the reading passage; a lexical comprehension section (LI), in which examinees suggest synonyms for four underlined words or phrases from the reading passage; a syntax section (SI), in which examinees complete sentences by using the syntactic modifications suggested in brackets; and, an essay for which several topics, based on the subject matter of the initial reading passage, are given (see table I).

The reading comprehension passage around which this particular ET was constructed, "Alternative to military service in Italy", seems to have been taken from a media report. It could be considered an appropriate text for low intermediate students, whereas an ET should be, if not an advanced test, at least a high intermediate test. There are no special difficulties as far as lexis, linking words, patterns or cohesion of the text is concerned. It is a very descriptive passage where the experience of a conscientious objector is presented through direct speech. The main obstacle for the testees is found in the first paragraph, where the group of objectors is categorised as "a large corps of community workers", whose function is described and where some less common core lexical items appear: "workers who have to shop for the disabled, tutor high school dropouts and take the elderly out". Although the facility / difficulty index [7] must be judged in relation to the students' proficiency level, this design should be evaluated in terms of how well it serves the utilitarian purpose for which it has been constructed. If it helps to spread students out into a normal distribution so that scores between performances are very different, the ET will fulfil the aims for which the test has been drawn up. If, on the contrary, there are items which do not accomplish the utilitarian purpose the ET has been built for and the results do not spread the scores out as expected, the validity of these items will be in jeopardy.

2.4. Scoring

Looking for homogeneity, the administrators of the entrance examination give rating cues for every item. It is assumed that with similar criteria there will be little room for disparity in the marks among raters. The following scoring scheme is provided for the raters together with some instructions, such as the relevant proportions to be assigned to comprehension, lexis, syntax and structure.

TABLE I

Scoring instructions

Item	Score	Competence	Type of the item	Technique
1.	0 - 2	communicative	Subjective	Open answer
2.	0 - 2	comprehension	Objective	True / False
3.	0 - 1	lexis	Objective	Matching
4.	0 - 2	syntax	Objective	Cloze
5.	0 - 3	communicative	Subjective	Non-directed essay

Evaluators are asked not to use more than two decimals in the final score. This suggestion leads evaluators either to use one decimal in marking each item or to use two whenever their feeling of accuracy invites them to do so and resort to the rounding system at the last moment on calculating the sum. From the outset the student knows the weight of each item in the final score as this appears in brackets following each question.

3. DATA ANALYSIS

3.1. Tables of frequencies

From a holistic perspective, the raters whose data were collected for this study use a similar scale of values. Whenever they do not mark with integers they resort to multiples of five: .25, .50, .75 ... etc., only one of the 8 raters also uses: .30, .80, 1.30, ... as alternatives. This particular interpretation could be explained on the basis of a need for rounding figures when the score of each item is added up. An illustration of the marking pattern drawn from the open item subtest data is offered in the following table: (Table II)

A tendency to centrality in the distribution of relative frequencies is observed. Raw scores on the OI subtest are spread out in an almost normal distribution. The mode is 1, the central value on the scale, the next value with the highest frequency is 1.5, and the lowest frequency level is found around the lower limit of the scale. A similar distribution has been obtained in the essay section. There is also a tendency to centrality. If scores within the range 0.5 - 2 are considered this distribution accounts for 65% of the frequencies. The mode and the median are slightly below the mean. Therefore, taking into account the information of the students' performance on both the OI and the essay subtests, the subjective ones, it would not be difficult to infer the skewness of the curve: slightly positively skewed for the essay and barely negatively skewed for the OI subtest. A quite useful form of information on

the behaviour of each subtest is available in Table III, where centrality and dispersion measures are given.

TABLE II

Open Item (OI)

		Absolute Frequency	Relative Frequency	Adjusted Frequency	Cumulative Frequency
Scale	.00	29	6.4	6.4	6.4
of	0.25	31	6.9	6.9	13.3
values	0.50	52	11.6	11.6	24.9
	0.75	26	5.8	5.7	30.7
	0.80	5	1.1	1.1	31.8
	1.00	80	17.8	17.8	49.6
	1.25	37	8.2	8.2	57.8
	1.30	1	0.2	0.2	58.0
	1.50	75	16.7	16.6	74.7
	1.75	44	9.8	9.8	84.4
	1.80	8	1.8	1.8	86.2
	2.00	62	13.8	13.8	100.0
	Total	450	100.0	100.0	

TABLE III

Statistics of the ET's subtests

		T/F	LEXIS	SYNTAX	OPEN	Essay
N	Valid cases	450.	450.	450.	450.	450.
	Missing cases	0.	0.	0.	0.	0.
Mean		1.8072	0.7582	1.3258	1.1393	1.3527
Median		2.0000	0.7500	1.5000	1.2500	1.2500
Mode		2.0000	1.00	2.00	1.00	1.00
Skewness		-2.4401	-0.839	-0.547	-0.248	0.243
Standard error		0.115	0.115	0.115	0.115	0.115
Mininum		0.00	0.00	0.00	0.00	0.00
Maximum		2.00	1.00	2.00	2.00	3.00

The distribution of the so-called objective subtests —True / False (TF), Lexis (LI) and Syntax (SI)— is quite different from the above-mentioned subjective items. The mean is below the mode and the median, which means that the curve is negatively skewed. The higher frequency of scores is found in the upper limit of the scale while few scores are found in the lower limit of the scale. The relative frequencies of the upper limit in the T/ F, LI and SI - 79.3%, 41,6% and 18.4%, respectively —are not balanced with their corresponding percentages read in the lower limit of the scale— 2%, 1,6% and 2.2%. In these subtests, the highest frequency is found in the maximum value of the scale.

There is no better way to illustrate the contrast between the scores for an objective and a subjective item than through a contingency table. The OI versus T/F pair is taken for this comparison because these two subtests average the same in the final score.

TABLE IV

Contingency table. Open item subtest vs True / False item subtest

		True / False (T/F)									
		.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	Total
Open	.00	1	2			5	3			18	29
item	0.25	1		1		6	1	1	2	19	31
(OI)	0.50	1				5	2	4	1	39	52
	0.75	2		1		1	1	2		19	26
	0.80							1		4	5
	1.00	2		1	1	7	1	4	1	63	80
	1.25	1				4		1		31	37
	1.30									1	1
	1.50					7	1	4		63	75
	1.75							4	1	39	44
	1.8							1		7	8
	2.00	1				4		3		54	62
Total		9	2	3	1	39	9	25	5	357	450

From a global point of view it is clearly shown that there is a very scant correspondence of frequencies in each value (leaving aside the detail of the .80, 1.30 or 1.80 values) in the rows and columns. A closer scrutiny of the table shows that if we merely consider all the frequencies which appear under the column with a 2 value of the T/F variable, then the distribution of the frequencies for each of the values of the OI variable would be: 18 and 54 in the lower and upper limits, a bimodal distribution with 63 frequencies in each case

and a relative tendency to centrality among the 357 students who scored 2 in the T/F pair. That is, the distribution we should have found in the other columns.

3.2. Shapes of distributions

If the distribution of a sample or population is normal, graphs are supposed to offer normal curves. In the sample studied such expectations are not fulfilled. Curves for each of the subtests are different from each other. An illustration of the curves obtained from the data studied can be seen in Figure I, a- e. There is some sort of parallelism in the subjective shapes which occur with enough regularity. This is something that cannot be said of the objective subtests items: the SI distribution could be seen as a cumulative graph while the LI and the T/F present leptokurtic and extremely negatively skewed curves.

3.3. T-tests

To compare the means and avoid the obstacle of different scoring scales (see Table I), all raw scores have been transformed into z-scores and taken to a scale from 0 to 10 (Table V.a). These transformations have allowed us to see how large the difference between the mean for the T/F subtest and that for the other subtests is, as shown in Table V.a.

TABLE V.a

Statistics for correlated samples

		Mean	N	S.D.	S.E.M.
Pair	LI-T	7.5816	450	2.5684	0.1211
1	T/F-T	9.0148	450	2.2150	0.1044
Pair	SI-T	6.6289	450	2.7390	0.1291
2	T/F-T	9.0148	450	2.2150	0.1044
Pair	OI-T	5.6967	450	3.0360	0.1431
3	T/F-T	9.0148	450	2.2150	0.1044
Pair	ES-T	4.5089	450	2.8944	0.1364
4	T/F-T	9.0148	450	2.2150	0.1044

Where S.D. means standard deviation and S.E.M. stands for standard error of the mean.

No less illustrative is table V.b. where pairs of correlations are established between T/F and all the subtests items. There is a weak and poor correlation, though significant because of the size of the sample.

TABLE V.b

Correlations of correlated samples

		N	Correlation	Sig.
Pair 1	LI-T vs T/F-T	450	0.245	.000
Pair 2	SI-T vs T/F-T	450	0.267	.000
Pair 3	OI-T vs T/F-T	450	0.209	.000
Pair 4	ES-T vs T/F-T	450	0.242	.000

Finally, these transformations have allowed us to apply the t-test for correlated samples. The data (Table V.c) show that in all the paired observations the critical values of "t" considerably exceed the critical value (1.96) for $p<.05$ in a two-tailed (non-directional) test. There are significant differences between the T/F subtest pair and the other subtests [8].

TABLE V.c

Test for correlated samples

		Differences					t	d.f.	Sig.
					Confidence intervals				
		Mean	S.D.	S.E.M.	Lower	Upper			
Pair 1	LI-T - T/F-T	-1.4332	2.9523	0.1392	-1.7067	-1.1597	-10.2976	449	.000
Pair 2	SI-T - T/F-T	-2.3859	3.0274	0.1427	-2.6664	-2.1054	-10.2976	449	.000
Pair 3	OI-T - T/F-T	-3.3181	3.3630	0.1585	-3.6296	-3.0065	-16.7183	449	.000
Pair 4	ES-T - T/F-T	-4.5059	3.1912	0.1504	-4.8015	-4.2102	-29.9526	449	.000

Where t refers to t-test, d.f. stands for degrees of freedom and Sig. means significance level.

A similar comparison can be established between each of the other objective subtests items and the rest (Table VI). Significant differences are also found in each pair for $p<.05$.

TABLE VI

Test for Correlated Samples

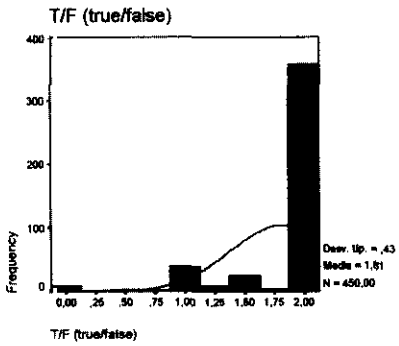
		Differences							
					Confidence intervals				
		Mean	S.D.	S.E.M.	Lower	Upper	t	d.f.	Sig.
Pair 1	LI-T - SI-T	0.9527	2.7493	0.1296	0.6980	1.2074	7.351	449	.000
Pair 2	LI-T - OI-T	-1.8849	3.1204	0.1471	1.5958	2.1740	12.814	449	.000
Pair 3	ES-t - LI-T	3.0727	3.0760	0.1450	-3.3577	-2.7877	-21.191	449	.000
Pair 4	OI-T - SI-T	-0.9322	2.9343	0.1383	-1.2040	-0.6603	0.6739	449	.000
Pair 5	ES-T - SI-T	-2.1199	2.7098	0.1277	-2.3710	-1.8689	-16.596	449	.000

4. DISCUSSION

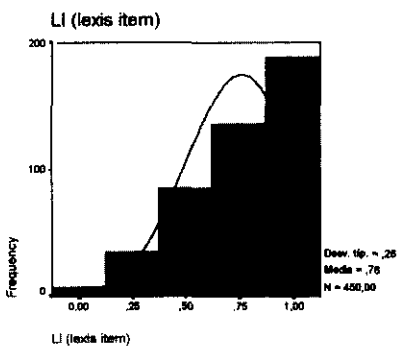
If the sample of a population is large and representative, the centrality measures, —mode, median and mean— are almost the same. In this study, it has been observed that there is a central tendency in the subjective subtest items (OI and essay) and a clear tendency to skewness in the objective subtest items /T/F, LI and SI) as can be seen in Figure 1, a-e. There must be some explanation for this behaviour in the latter subtests, and this must be found more in internal than in external circumstances. If students and raters are the same, there is no reason to focus the attention on them because they write and mark the subjective items in the same way as the objective ones. Thus, attention must be focused on the item as such and the facility of its accessibility as the cause for this degree of skewness of the curves. If the distribution of the subjective items (Fig.1.d and e) is quite acceptable, the values of the skewness in the objective ones (Fig.1. a, b, and c) is so highly negative that it leads one to interpret the facility / difficulty index as the main reason for the high scoring in these items (Table III). Not only following Feldt's index (1993) but also admitting Fulcher's broad index (1997), the T/F and the other objective items will be inappropriate due to their weak power of discrimination. As testing techniques, T/F, LI and SI are acceptable but the way they are presented fails. A similar conclusion can be reached taking the 25 and 75 percentiles: 0,50 and 1 for LI , 1 and 1,75 for SI and 2 and 2 for T/F. That is, if all the data were taken to a relative cumulative frequency curve in the first quartile, the 25th percentile, we would find that only 25% of students fail to reach (.50 and 1) in the LI and SI subtests, respectively. The percentile decreases in those which fail to score the maximum value of the scale in the T/F subtest. A glance at the relative cumulative frequency curve shows that the third quartile, the maximum value of the scale for LI

FIG. I (a-)e

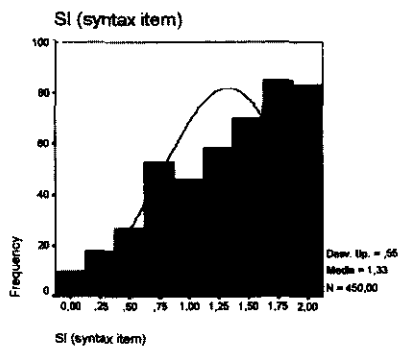
a.



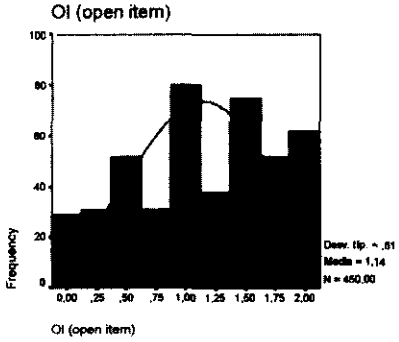
b.



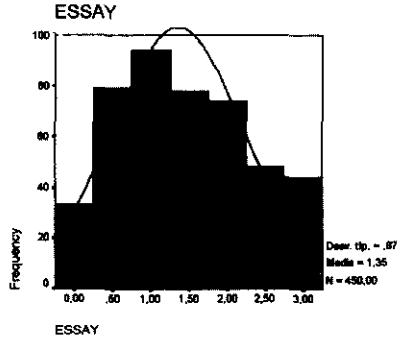
c.



d.



e.



and T/F subtests is already found, whereas in the SI subtest fail to reach 1.75 out of 2.

As the T/F item is the focus of most of this analysis, a specific discussion on the distribution of frequencies and of the scale of values used by the evaluators is required. In the former issue it is supposed that, if the marking instructions of the University Examination Board [9] had been followed, the frequencies in a normal distribution should have grouped at "1" if one answer is correct, at "2" if both answers are correct or at "0" if both are incorrect. Had the item been properly calibrated, a good balance between the degree of difficulty and the proficiency level of the skill, the frequencies would have mainly grouped at "1". This value would have been the converging point for the mode, median and mean. But in the data obtained "2" appears as mode, median and almost as mean since the value of this statistic is 1.8072. Thus, it can be stated that the T/F item has not been properly calibrated when it was designed and that its facility index is so high that it can be considered a redundant item with an insignificant discriminative power.

A thorough revision of the scale of values shows the subjective factor when scoring. If the mentioned instructions had been taken into account, there would not have been room for personal interpretations. It is an objective item with a scale of three values. Nevertheless, the frequencies found under columns other than those of "0", "1" and "2" amount to 45, as can be read in the contingency table (Table IV). This finding shows that a degree of subjectivity bias encroaches upon an objective test item at the time of scoring.

Similar comments can be stated in regard to the other objective items (Table III), lexis and syntax. Their skewness, their significant differences and their subjectivity bias when they were marked lead to the same conclusion. According to the Classical Test Theory (Crocker and Algina 1986), these items have been inappropriately and badly calibrated for this ET. These findings provide enough arguments to propose a re-examination of the marking system and to call for a debate among raters and academic authorities seeking homogeneity of criteria.

Little more can be added to the flagrant disparity of the distribution of frequencies of this item if it is not contrasted with the distribution of another item. The reading of Table IV gives a new insight into the issue. To highlight the contrast of the distribution, the contingency table is drawn on OI and the T/F. As can be seen in figure 1. d.e., the distribution of frequencies in the subjective items can be taken as almost normal whereas the frequencies on the objective items are not spread out as should be expected. If the correlation between these items had been close to "1", things would have been different, but its value (.242), though significant, is not strong. Assuming that the subjective OI subtest has an acceptable discriminating power, 54 students approximately should have got the same mark on the objective T/F subtest.

Deducting these 54 students, who scored 2 on both subtests, from the 357 people in the column under the highest mark of T/F it will be found that about three hundred students have benefited from the design. If the 45 scores attributable to the subjectivity bias are added to this figure, it will be found that as many as 350 students out of 450 have been rewarded for one or another reason on the T/F items. From the perspective of those who have got the maximum score on the subjective subtest item, they can consider themselves to have suffered some sort of penalisation. They would probably still have got the maximum score had the objective item been more highly discriminative.

The contingency table has led us to focus on the contrast between T/F and OI, but similar pairs of contrast could have been set up between T/F and Essay, or between LI and any of the subjective items. The inferences would have been similar since the discriminative power of Lexis is almost as weak as T/F.

So far, these data allow us to call for a revision of the objective items. It is open to argument whether the objective items are representative or not of the morphosyntatic domain tasks in the structure or framework of a ST, but, in this study, the nature of these items themselves, their facility / difficulty index, is questioned. They do not fulfil the utilitarian purpose they were built for: to rank students entering the university, that is, to spread their scores out into a normal distribution. The subtest items accord neither with teleological theories nor with deontological theories (Davies 1997). If the former are primarily concerned with validation in terms of outcome the principal focus of the latter is fairness. The items being questioned here are, on the one hand, inherently inappropriate. Students who have attained a higher standard of English have been penalised in favour of those students whose standard is lower. On the other hand, these test items do not bring about the best results: an accurate discrimination.

The highly negative skewness of the distribution of frequencies for the objective test items allows for the application of t-tests to related samples. The results, as seen in Tables V and VI, confirm the hypothesis that there are significant differences between the objective and subjective items. The objective ones average higher than they are supposed to, their highly negative skewness and consequently their low level of discrimination question the validity of the ST. Theoretically, they were supposed to account for the same level of discrimination and to average approximately the same in the final score, but in practice they discriminate less than the subjective ones and the weight of the objective test items in the final score is higher. Most of the significant differences between the T/F and the rest of the items are mainly explained on the grounds of the design and to some extent, they may be due to the subjectivity bias, though this latter factor could affect all scores in the same

way. This bias has been uncovered because of the Examination Board's precise instructions in a few cases, but it can be taken for granted that a dose of hidden bias operates, though in a random way, on the remaining scores. It can be argued that other factors such as partial knowledge, guessing effect, strategies to close the gap and a whole range of other test-taking behaviours could play a part in the data obtained. Nevertheless, a thorough examination of each of these items shows that such issues affect everybody in the same way. Hence, they can not contaminate the outcome of this study.

5. CONCLUSION

This study provides enough evidence to state the following claims;

— The highly negative skewness, ceiling effect, in the objective items (T/F, LI, and SI) shows that the scores are not spread out into a normal distribution.
— There are significant differences between the objective and subjective items(OI and essay).
— Lack of calibration in the design of the objective items is evident.

That this study shows such claims to be well founded leads to the following conclusions:

1. The objective items, which average 50% of the final scores, affect the goal of the test: to discriminate among students. Consequently, the discrimination of the students' performance merely rests on the subjective items.
2. The distribution of the scores induces us to think that the objective items are nearer to a criteria-referenced test, a minimal competence test, than to a norm-referenced one, when what really matters is the score of one student in relation to the others.
3. The ET is in conflict not only with the teleological theories but also with the deontological theories. The former are not upheld because the utilitarian purpose the test was built for has not worked out properly. The latter cannot be sustained in the face of the test's lack of fairness: better students have been penalised in favour of those whose knowledge of the language was poor.
4. The ET paper has not averaged in the final score of the Spanish university entrance examination in the way it was supposed to. Consequently, there are students who have obtained a final score above their merits while students who might have deserved a better mark relative to others have not got it. This situation could give rise to significant, even dramatic personal consequences: one student may have been unfairly deprived of the decisive decimal points to gain entry to the faculty of his / her choice

while another student may have been unfairly awarded that of his / her choice.

5. Hence, it can be concluded that the validity of the objective items is called into question in the ET studied and that attention should be focused on the design and calibration of the objective items in order to guarantee the validity and the discriminative power this specific English Proficiency Test should have.

ABBREVIATIONS

ELI: English Language Institutes
CLA: Communicative language ability
COU: University Orientation Course
ET: English Test
LI: Lexis items
OI: Open items
SI: Syntax items
T / F: True / False
LI- T. Lexis item transformed. ...

NOTES

[1] This research has its origin in the project "Analysis on test's parameters" (ref: PR94-118), carried out at the Measurement and Computer Analysis Department in OISE, Toronto, Canada, with the financial support provided by the DGICYT. My acknowledgements also to Mª Rosario Martinez Arias and Michael White for their helpful comments on the draft and to the anonymous referees for their patience and interest.

[2] The use of ET henceforth will refer to the English Test in the Spanish University Entrance Examinations

[3] In a placement test, testers may also be concerned with how well students will learn if they are placed in a given group or if a particular sequence of language course objectives has been fulfilled (Cumming and Berwick, 1995). Our ET is not meant to allocate students in future English classrooms. It is not designed to find out what they know but rather its purpose is to show how well they perform in relation to the others. It is taken in the Spanish University Entrance Examination as a subject which contributes to average the final score in the same way as Mathematics, Philosophy or History do. Our academic authorities, our society, —both institutions and students— demand a reliable ranking in the final score.

[4] COU refers to the studies taken in the year prior to entering the Spanish University.

[5] The data have been studied with the standard statistical package for social sciences. 8.01. As it is a Spanish version an English translation has been provided.

[6] For the sake of representativeness, each rater was asked to transcribe the first thirty marks of every center corrected, irrespective of the fact that the number of students from the different centres varied considerably.

[7] Feldt considers an item to be appropriate if the index is close to .5%, while Fulcher (1997) proposes a range of acceptability between .30 - .70, a range previously used in Herrera (1996).

[8] Technical help for the interpretation of the data presented, if required, can be found in Woods, Fletcher and Hughes (1986), and in Butler (1985).

[9] In this question the student must first answer true or false and secondly he must give evidence for his/her answer quoting the text on which he / she bases the answer. The score for each question in this item will mean 1 point. *The score will be zero points if the correct evidence is not given;* in relation to this issue an incomplete quotation or just the pointing out of the line/ lines will not be accepted. A contradiction between the quotation and the truthfulness or falsehood of the answer will also be scored as zero.

Answers:

a. True. "The number of young Italian men who avoid military service by stating they are conscientious objectors has risen sharply in recent years."

b. False. "The army would be a lot easier. They give you an order and you follow it".

Departamento de Filología Inglesa
Facultad de Filología
Universidad Complutense de Madrid

REFERENCES

Alderson, J.C. (1991). Language testing in the 1990s: how far have we come? How much further have we to go? in Anivan, S., (ed.), *Current Developments in Language Testing,* Singapore: Regional Language Center, 1-26.

Bachman, L.F. (1990a). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Brown, J.D. (1988). *Understanding Research in Second Language Learning.* Cambridge: Cambridge University Press.

Butler, C. (1985). *Statistics in Linguistics.* Oxford: Basil Blackwell.

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1: 1-47.

Canale, M. (1983). On some dimensions of language proficiency, in Oller J.W.jr.(ed.). *Issues in language testing research.* Rowley , MA: Newbury House, 333-42.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing* 14: 3-22.

Chastain, K. (1988). The ACTFL proficiency guidelines: a selected sample of opinions. *ADFL Bulletin* 20: 47-51.

Crocker, L. and Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* Chicago, IL: Holt, Rinehart & Winston.

Cumming, A. and. Berwick, R. (1995). *Validation in Language Testing.* Clevedon: Multilingual Matters Ltd.

Davies, A. (1997). Introduction: the limits of ethics in language testing. *Language Testing* 14: 235-241.

Feldt, L.S. (1993). The relationship between the distribution of item difficulties and test reliability. *Applied Measurement in Education,* 6: 37-48.

Fulcher, G. (1997). An English language placement test: issues in reliability and validity. *Langage Testing* 14: 113-138.

Harlow, L. and Caminero, M. (1990). Oral testing of beginning language students at large universities: is it worth the trouble? *Foreign Language Annals* 23: 489-501.

Herrera Soler. H. (1996). Implicaciones metodológicas de una elección múltiple. Barrueco, S., Hernández, y L.Sierra, (eds.) *Lenguas para Fines Específicos IV:* 469-475.

Kenyon, D.M. and Stansfield, C.W. (1992). Examining the validity of a scale used in performance assessment from many angles using the many facet Rasch model. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA (ERIC Document Reproduction Service, ED 343 442).

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing,* 7: 52-76.

Wall, D.; Clapham, C. and Alderson, J.C. (1994). Evaluating a placement test. *Language Testing* 11: 321-344.

Woods, A.; Fletcher, P. and Arthur, H. (1986). *Statistics in Language Studies.* Cambridge: Cambridge Textbooks in Linguistics.