Estudios Ingleses de la Universidad Complutense ISSN: 1133-0392
Vol. 10 (2002) 113-137

A new insight into examinee behaviour in a multiple-choice test: a quantitative approach

Honesto Herrera Soler
Rosario Martínez Arias

Universidad Complutense de Madrid

ABSTRACT

In the context of both the facility / difficulty index and the examinees' ability in a multiple-choice item, this study is an attempt to gain a deeper view of these issues in a test targeted to a personnel selection process. Our analysis compares the data supplied by the Classical Test Theory (CTT) and the Item Response Theory (IRT), and follows examinee behaviour on the distractors in the hope that the information obtained through these theories will give us a better insight not only into the item and examinee ability but also into the role the distractors may play in this sort of test. On the basis of the information provided we claim that if the test was intended to measure a desired level for a personnel selection process, the test designer has not succeeded in providing even a single item where the correct option and appropriate distractors fulfill the expected criteria of a good item, at least for upper intermediate level candidates. Since the degree of difficulty is a matter not only of the item but also of the distractors, the test should be thoroughly revised and many distractors modified.

Key words: Classical Test Theory (CTT), Item Response Theory (IRT), fractile, ability, difficulty index, slope, threshold, asymptote, biserial correlation, formula scoring.

RESUMEN

UNA NUEVA APROXIMACIÓN AL ANÁLISIS DE LA RESPUESTA AL ÍTEM EN UN TEST DE ELECCIÓN MÚLTIPLE: UNA PERSPECTIVA CUANTITATIVA

Este trabajo es un intento de profundizar en los problemas que generan los índices de facilidad / dificultad de los ítems y el conocimiento que se tiene de los mismos en un test de elección múltiple aplicado a un proceso de selección de personal. Nuestro análisis compara los datos obtenidos tanto con la Teoría Clásica de los Tests como con la Teoría de la Respuesta al Ítem y estudia el comportamiento ante los elementos de disuasión del ítem de quien realiza el test, con la esperanza de que la información obtenida nos facilite no sólo la valoración del ítem y del nivel del sujeto sino también el papel que los elementos de disuasión pueden jugar en este tipo de tests. A partir de la información proporcionada por ambas teorías consideramos que si el test se diseñó para evaluar un determinado nivel en un proceso de selección de personal no se ha conseguido ni siquiera un solo ítem en el que la opción correcta o los elementos de disuasión cumplan los criterios que se esperan, al menos en el caso de unos candidatos de un nivel medio alto. Dado que el índice de dificultad depende no sólo del ítem sino de los elementos de disuasión, tanto uno como otros deberían revisarse.

Palabras clave: Teoría Clásica de los Tests, Teoría de Respuesta al Ítem, fractil, conocimiento, índice de dificultad, índice de discriminación, índice de respuesta al azar, correlación biserial.

1. INTRODUCTION

Until quite recently these sorts of studies have traditionally been carried out resorting to the Classical Test Theory (henceforth CTT), but the impact of the Item Response Theory, henceforth (IRT)[1] in the 20th century has led researchers to look for new models. Skehan (1989:3) discusses its influence on recent language testing research: "One area where considerable progress has been made … is that of statistical techniques, for both reliability assessment as well as for validation. With the former, the most noteworthy development has been the extensive applications of item response theory (or latent trait measurement) to language testing."

1.1. CTT and IRT

In the Classical Test Theory (CTT), we assess the difficulty of an item by its *p-value*, that is, the proportion of people in the sample who have responded correctly to the item. The higher the p-value, the easier the item, i.e. 0.9 (90%) of correct answers will tell us that the item is extremely easy. Therefore, this theory depends on the low or high abilities of the sample studied. If the examinees' level is above the item difficulty the p-value will be very high, or the contrary if their level is low.

Efforts were made among researchers to overcome this sort of dependency throughout the last century and work was oriented towards the development of measurement procedures in which the scores on the test were not testdependent and could be standardised across similar tests despite different ability levels. That is what precisely we obtain with Rasch's model (1966), one among other IRT models. Rasch's model is described as the one-parameter model versus other IRT models known as two- and three-parameter models. At its most fundamental level, the Rasch model assumes that a response to an item is a function only of the student's ability and the item's difficulty (Ludlow and O'Leary, 1999:619). It is a theory that provides a description of the relationship between an examinee's ability level on the construct being measured by the item and the probability that the examinee will respond to the item correctly.

The advantage of using Rasch's model over other IRT models lies in the model's assumption of the separability of the parameters of examinee ability and item difficulty (Griffin, 1985: 151). This is something that the CTT has not overcome, since, as we have said before, in this theory the individual's score is a function of the item difficulty and the person's ability. With the Rasch model the probability of a candidate achieving a correct response is purely a question of the difference between the candidates's ability and the difficulty of the item. This property, unique to the Rasch model, allows the comparison of two stimuli independently of which particular individuals are instrumental for the comparison and at the same time a comparison between two individuals independently of which particular stimuli are instrumental for the comparison (Wilson 1991).

Both CTT and IRT are appropriate in the exploration of test constructs, though the information provided by the IRT is more exhaustive. The latter allows for the comparison of an individual's ability in different tests, both item and individual ability being measured on the same scale and it may also become an important tool in the building of computerized adapted tests (CAT). Since our sample is large enough for the one-parameter model and it is not large enough for the two-parameter and three-parameter models, we will focus our research first on finding out the information provided by the CTT and IRT when the ability / difficulty dimension of each item of the test is assessed, and then on the candidates' behaviour in regard to distractors in the different fractiles[2].

1.2. Contribution of a quantitative approach to the knowledge of individual items

If we know the quality of each item in a test, we should be able to deduce the quality of the total test score (Hoi 1990). This viewpoint has led us to consider the examinee's linguistic performance in each item of the test studied. This information will provide a quantitative approach to the examinee's behaviour and a powerful insight into some of the characteristics of individual items that will allow the test designers to improve the weak items.

For a multiple choice test such as the one we are analysing, an inherent phenomenon is the probability that item score is affected by guessing. Unfortunately, that behaviour is not well understood today (Lord 1975, Bliss 1980, Fulcher 1997, Herrera 1999)[3]. We generally asssume that people lacking the necessary knowledge to select the correct answer would guess at random. Based on the random guessing model, for a multiple-choice item with *m* options, the probability that a subject can answer an item correctly through guessing is *1/m,* i.e. 0.25 (25%) when the item presents four options.

There are two views regarding the probability of a correct guess. One view suggests that people who do not know the correct answer generally have some partial knowledge (Crocker and Algina 1986), thus they are able to eliminate some distractors. Therefore the probability of choosing the correct response increases from 25% to 50% if two of the three distractors are eliminated on a 4-option multiple choice item. An alternative view goes in an opposite direction. Item writers tend to generate distractors that are not only plausible but attractive (Lord 1974). This design leads to the idea that examinees lacking the necessary knowledge to choose the correct option would be attracted to the incorrect options. Hence, the probability of a correct guess is lower than *1/m*. In both cases in order to maximise the ability of the observed score representing the true score, the effects of guessing should be removed from the observed score through the formula scoring[4] (Lord 1975).

In a design of this sort of test it is taken for granted that test writers have made an effort to find not only plausible but also attractive distractors. From this perspective it is assumed that the proportion of correct answers in a wide interpretation is within the 0.3 (30%) to 0.7 (70%) range and it should be expected that the effects of guessing in addition to the attractiveness of the distractors has meant that each option has been chosen by about 10% of the examinees, that is, the 0.3 (0.30%) left. On this basis, an ideal and perfect model of answers to a multiple choice item should contain the figures just mentioned, an aim that is considered to be unattainable in practice. Thus, it is necessary to resort to "the goodness of fit models" in Statistics, which helps us to check the degree to which a mathematical model or theoretical distribution fits a set of observed data. Although we are aware that we will not come across ideal distributions in the choice of options we will take them as a starting point to see to what extent the distribution observed in each item of the test under study approaches our theoretical model.

2. METHODOLOGY

2.1. Skill and testees

The test under study deals with reading, one of the skills that any test intended to assess the ability of candidates to communicate effectively in their workplace should have. It is a test targeted to a personnel selection process, and has been piloted with students taking English as a foreign language in ICADE, in the Faculty of Economics at the Universidad Complutense de Madrid, all of them familiar with the business and economics register, and also with students of the Escuela Oficial de Idiomas more concerned with general English. As a result of their background it is assumed that they have an upper intermediate level[5].

Large samples are required in most of the IRT models, but the one-parameter Rasch model is not so demanding and the size required is widely fulfilled in our case, since the test studied has 60 items and 321 subjects for analysis[6]. These figures also allow us, to some extent, to work with the two- and three- parameter model (Hambleton *et al.* 1991, and De Jong and Stoyanova 1994).

Reading comprehension research has shown that textual background knowledge (content schemata) and text structure knowledge (formal schemata) can vary from one language to another and from one student to another in accordance with their knowledge background (Curtis & Glaser, 1983). Thus, to control these variables and avoid biased interpretations we have chosen to control text property by studying item responses only from Spanish examinees involved in undergraduate studies.

2.2. Material

Content: The items could be considered to lie within the core of general English although the register was that of business domain.

Format: Four 15-items series were prepared on the following structural basis. In the first, candidates are required to find among the options presented, the one that best fits the proposition offered in the prompt. In the second series testees are asked to fill in the gap for an item with the appropriate term, each item having its own context. In the third series candidates must cope with questions and answers. In the last one candidates go back to a fill-in design as in the second series, but in this case within a contextual situation. The instructions given were the following:

a. Choose the sentence that means the same as the one given.
b. Circle the most appropriate option for each sentence.
c. Choose the sentence that most accurately answers the question.
d. Circle the most suitable option to fill in each blank.

2.3. Administration circumstances

Candidates were provided with a sheet for the answers together with the test. They were allowed 40 minutes to complete the test. The test was administered during a normal teaching period.

2.4. Psychometric Analysis

Classical test theory analyses were carried out with the SPSS 10.01 and the TESTFACT program, where fractiles are taken as quartiles for each item in our study (Wilson, Wood, Downs and Gibbons, 1991). Rasch model analyses (dichotomous) were realised with the BILOG (Mislevy and Bock 1989).

3. RESULTS

In the Spring of 2000, 321 candidates attempted the reading test, 236 completed the test and 85 were not able to complete it. We decided to take all the candidates into account whether they had finished or not. Nevertheless, we distinguished between omitted item response – those in which a student skips an item by mistake or reads an item and consciously decides not to answer it – and a not-reached response which may occur when the student does not have the opportunity to answer an item, usually because of lack of time. Figures for omitted item response were low while those labelled as not-reached response were considerable. As there are usually problems in fitting long tables onto a page we focus our comparison on the quantitative information provided by CTT and IRT on the first 40 items, where there are just a few examinees in the not-reached response category.

3.1. CTT and IRT contribution

The following table provides a summary of selected statistics from the CTT perspective.

Table 1
CTT

Item	*Subjects*	*Number right*	*Facility index*	*Biserial correlation*
1	321	244	.760	0.090
2	321	253	.788	0.451
3	321	289	.900	0.453
4	321	124	.386	0.044
5	321	272	.847	0.515
6	321	178	.963	0.834
7	321	309	.963	0.834
8	321	300	.935	0.487
9	321	234	.729	0.488
10	321	292	.910	0.459
11	321	239	.745	0.561
12	321	237	.738	0.499
13	321	254	.791	0.386
14	321	170	.530	0.397
15	321	190	.592	0.399
16	321	299	.931	0.447
17	321	237	.738	0.661
18	321	195	.607	–0.011
19	321	305	.950	0.432
20	321	211	.657	0.558
21	321	257	.801	0.715

22	321	298	.928	0.740
23	321	302	.941	0.846
24	321	286	.891	0.721
25	321	151	.470	0.539
26	321	278	.866	0.853
27	321	265	.826	0.624
28	321	302	.941	0.734
29	321	308	.960	0.632
30	321	295	.919	0.579
31	321	299	.931	0.579
32	321	283	.882	0.690
33	321	267	.832	0.570
34	321	246	.766	0.560
35	321	284	.885	0.692
36	321	61	.190	–0.141
37	321	255	.794	0.756
38	321	267	.832	0.612
39	321	41	.129	0.103
40	321	205	.639	0.582

It includes the following data in its output: Number of candidates, right answers, facility / difficulty index, and item discrimination and the biserial correlation. Our analysis from the CTT takes into account both the Facility / Difficulty Index and the homogeneity index computed by Biserial Correlation. The referent criteria in the Facility / Difficulty index should be within a range of 0.30 - 0.70. Below or above these limits they are considered extremely difficult or easy respectively. We work with the data provided by the biserial correlation, which is less sensitive to extreme values than the Pearson correlation. The criterion reference to see if an item shows consistency with the total scale of the test is >0.25.

A reading of the different values under each heading shows that in spite of the different scales of measurement a clear correspondence is observed on the way the items behave. If an item is labelled as easy or anomalous, it can mostly be taken as such on the different scales. Let us take items: 4, 18, 36 and 39. They are anomalous in most of the parameters. They are outside the conventional range of the facility index that we mentioned before. The biserial correlation is not within standard parameters: item 36 would fall together with item 18 within the category of the absurd items since they have negative values. The data found in items 4 and 39 also fail to contribute to the consistency of the test.

On the whole, the facility index in most of the items is above the range commented. As in the CTT the facility /difficulty index depends on the ability level, we reach the conclusion that in this piloting the items presented were too easy for the candidates taking the test.

Table 2
ITEM RESPONSE THEORY

Item	*Subjects*	*No. right*	*Slope*	*Threshold*	*Asymptote*	*Chisq**
1	321	244	0.401	–1.684	0.292	7.0
2	321	253	0.914	–1.139	0.250	6.2
3	321	289	0.773	–2.730	0.251	3.5
4	321	124	1.053	2.612	0.320	6.4
5	321	272	1.170	–1.261	0.334	13
6	321	178	1.421	0.516	0.275	4.6
7	321	309	1.510	–2.778	0.229	1.4
8	321	300	0.844	–3.903	0.243	7.1
9	321	234	0.913	–0.798	0.224	4.8
10	321	292	0.881	–2.610	0.253	2.5
11	321	239	1.570	–0.391	0.323	4.6
12	321	237	1.243	–0.488	0.300	4.2
13	321	254	0.789	–1.329	0.261	7.0
14	321	170	0.997	0.519	0.216	10.0
15	321	190	1.010	0.268	0.254	6.7

16	321	299	0.799	–3.281	0.241	2.4
17	321	237	1.684	–0.523	0.246	3.7
18	321	195	0.535	2.622	0.483	7.8
19	321	305	0.806	–3.722	0.243	0.5
20	321	211	1.564	–0.083	0.259	9.1
21	321	257	1.874	–0.774	0.278	5.3
22	321	298	1.382	–2.290	0.210	7.1
23	321	302	1.528	–2.330	0.225	1.3
24	321	286	1.413	–1.752	0.233	1.8
25	321	151	1.507	0.497	0.150	8.9
26	321	278	1.848	–1.393	0.1950	1.2
27	321	265	1.179	–1.389	0.205	5.4
28	321	302	1.376	–2.439	0.235	3.0
29	321	308	1.105	–3.185	0.246	1.0
30	321	295	1.022	–2.561	0.226	1.8
31	321	299	0.69	–2.867	0.231	10.2
32	321	283	1.317	–1.770	0.212	6.2
33	321	267	0.972	–1.636	0.208	11.7
34	321	246	1.129	–0.918	0.229	8.3
35	321	284	1.228	–1.890	0.208	6.5
36	321	61	0.517	—	0.250	21.2
37	321	255	1.616	–0.955	0.202	5.5
38	321	267	1.103	–1.508	0.213	12.3
39	321	41	0.998	—	0.244	25.6
40	321	205	1.481	–0.141	0.204	4.8

There are four points of interest in Table 2: slope, threshold, asymptote and χ^2. Through the slope column we learn the discriminating value of the item. The threshold column shows the degree of difficulty of each item in a different scale from that of the CTT scale that runs from *0 to 1*. In the IRT a standardized metric is assumed (mean = 0 and SD = 1) and items are considered more or less difficult depending on the distance from the midpoint of the curve. It is understood that the further left the score of the midpoint the easier the item is and, on the other hand, the further right, the more difficult. The standard criterion for the guessing effect, the asymptote column, in this format of multiple choice is around 0.25. The fit model, that is, the relationship between the observed and the expected frequency, is analysed through the χ^2, where there is not a significant difference in any item with p-value <0.1

3.2. Information provided by the fractiles

At first glance, when considering the data in Tables 1 and 2, Facility / Difficulty index, we realised that there was no single item that fitted the ideal paradigm above mentioned, that is, a range of 30-70 % of the choices for the right option and about 10% for each distractor. The data obtained through the fractiles showed how far they are from this ideal paradigm. It was observed that the percentages for the right answers in a considerable number of items were above 80%, whereas percentages within a range of 0% and 5% were found in several distractors.

This information encouraged us to develop a new paradigm to reflect the candidates' behaviour in their answers. It can be observed that in quite a number of items one of the options was chosen by the majority of the examinees and it happened to be the correct one. Items inviting this performance are labelled a *dominant option*. There was also a significant number of answers where the examinees were able to leave out two of the wrong answers, so the multiple choice format turned out to be a *true / false test*, with a 0.5 (50%) *p-value*. And finally, a third category, labelled as *anomalous* was identified. The trait that defines this category is that one of the distractors presents as many or more frequencies than the correct option. Thus, on the ground of this information we present the following category paradigm (fig. 1).

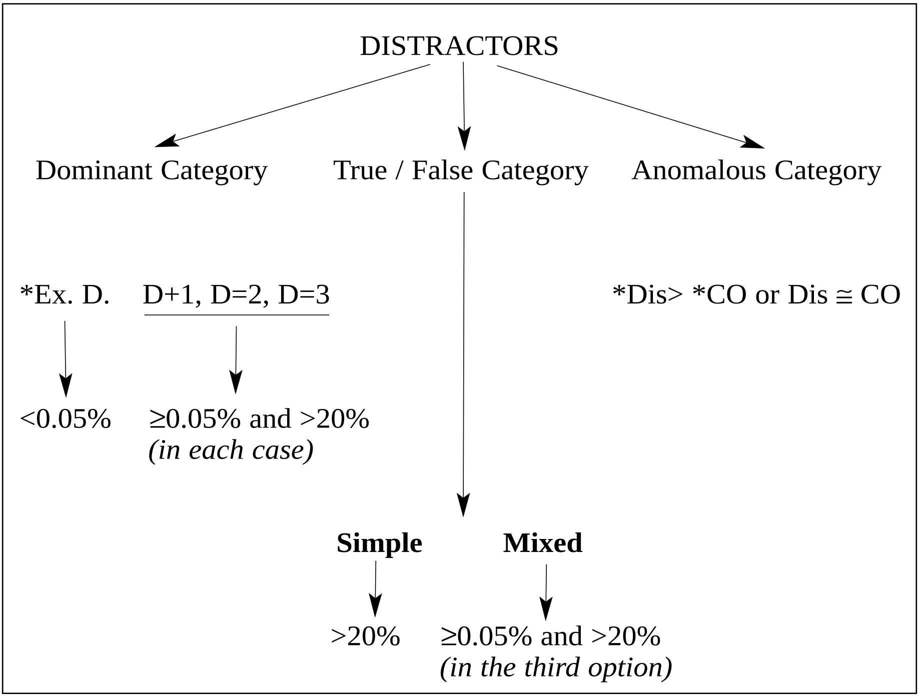

Fig. 1. Distribution of distractors.
Where *Ex D. stands for exclusively dominant; *Dis for Distractor and *Co for Correct.

The *dominant category* consists of those items where the chosen option is the correct one and the other three added together are favoured by less than 20% of the candidates. Within the dominant category we defined four different subclasses:

a. Exclusively dominant.
b. Dominant +1.
c. Dominant +2.
d. Dominant +3.

Exclusively dominant means that the chosen option is the correct one and the other three are virtually not chosen at all. *Dominant +1* arises when the chosen option is the correct one, and one of the other three is chosen by less than 20% of the subjects. The third subclass, *dominant +2,* indicates that the chosen option is the correct one and two of the distractors are chosen by at least 5 % of the subjects and the 4th is practically rejected outright. Finally, *dominant +3* expresses the idea that the chosen option is the correct one and the other three distractors are selected by at least 5% of the students.

The *true-false category* is made up of those items where the correct option and one of the distractors are chosen by at least 20% of the examinees. It is subdivided into:

a. *Simple True / False category*: two options including the correct one are chosen by at least 20% of the candidates and the other two are selected by less than 5% of the candidates.

b. *Mixed True / False category*: two options including the correct one are chosen by at least 20% of the candidates and the third option is followed by at least 5% of the candidates.

Finally, the *anomalous category* comprises those items where the highest frequency is given to an incorrect.

The paradigm presented is better illustrated in the fractiles tables 3 and 4, where prototypical examples of each category are offered:

Table 3 is a model of the complete information provided by fractile analyses. There is a first part where for each score band number, percentage and means in the totals and omissions are presented. The data of the second part, responses, are oriented to analyse the examinees' behaviour in each item. As we are interested in the latter we offer Table 4, where prototypical items in each category of the paradigm are presented. If item "29" is labelled as *exclusively dominant* owing to the low percentages: 0.6 for a *not-reached response* under code: 8 and: 2.2; 0.0; 0.6 for the distractors, the others are labelled as dominant +1, +2, or +3 in accordance with the percentages reached in the different distractors. Together with the relative frequency under each distractor, the whole-test performance *means* of the candidates choosing this distractor have been supplied.

If we read not only frequencies but also percentages under each option (table 4) we will find the linguistic behaviour of the candidates. There will be a progressive increase in percentages in the correct option as we move from the first to the last fractile and a decrease in the case of the distractors in all items except in those labelled anomalous, in which there is either no discrimination or an inverted order in the different fractiles.

From the fractile perspective items are allocated in the following way. Within the so-called *dominant category* we found 47 items out of a total of 60, which represents 78.3%. The *"true-false category* covers nine items or 15% of the total and in the *anomalous* one we found four items or 6.6%.

Table 3
RESPONSE BY FRACTYLES

ITEM 29. EXCLUSIVELY DOMINANT

FRACTILE		*TOTAL*		*TOTAL+OMIT*		*OMIT*
SCORE BANDS		*N*	*%*	*N*	*%*	*N*
0	36	77	24.21	78	24.5	1
37	45	79	24.8	79	24.8	0
46	51	86	27.0	86	27.0	0
52	60	76	23.9	76	23.8	0
TOTAL	*N*	318		319		1
	%	99.7		100		0,3
	$\bar{X}$	99.7		100		0,3

RESPONSES

FRACTILE SCORE BANDS		“8”	1*	2	3	4
0	36	2	69	5	0	1
38	45	0	76	2	0	1
47	51	0	86	0	0	0
52	60	0	76	0	0	0
TOTAL	*N*	2	307	7	0	2
	%	0.6	96.2	2.2	0.0	0.6
	$\bar{X}$	13.5	43.6	32.3	0.0	0.6

Where: * = correct option
“8” = not-reached responses
N = number of candidates
% = percentage
X = Mean in the test of the subjects belonging to that group

Table 4
RESPONSES TO SOME PROTOTYPICAL ITEMS

ITEM 26. DOMINANT + 1

FRACTILE SCORE BANDS		“8”	1*	2	3	4
0	36	2	44	2	9	18
37	45	0	73	0	2	4
46	51	0	85	0	0	1
52	60	0	75	0	0	1
TOTAL	*%*	0.6	86.8	0.6	3.4	7.5
	$\bar{X}$	13.5	45.3	28.3	22.2	30.5

ITEM 11. DOMINANT + 2

FRACTILE SCORE BANDS		“8”	1	2*	3	4
0	36	0	4	34	8	29
37	45	0	1	59	5	13
46	51	0	0	71	3	11
52	60	0	0	75	0	1
TOTAL	*%*	0	1.6	74.9	5.0	16.9
	$\bar{X}$	0	29	45.7	35.3	36.0

ITEM 15. DOMINANT + 3

FRACTILE SCORE BANDS		“8”	1	2	3	4*
0	36	0	21	16	7	30
37	45	0	16	14	5	39
46	51	0	10	15	4	55
52	60	0	3	5	2	66
TOTAL	*%*	0.0	15.7	15.7	5.6	59.6
	$\bar{X}$	0.0	37.3	41.1	38.7	46.1

ITEM 44. SINGLE TRUE/FALSE

FRACTILE SCORE BANDS		“8”	1	2	3*	4
0	36	30	25	4	12	3
37	45	0	41	2	33	2
46	51	0	28	1	55	1
52	60	0	12	0	63	1
TOTAL	*%*	9.4	33.2	2.2	51.1	2.2
	$\bar{X}$	24.1	42.1	37.3	48.1	39.1

ITEM 52. MIXED TRUE/FALSE

FRACTILE SCORE BANDS		"8"	1	2	3*	4
0	36	49	11	10	3	4
37	45	11	29	14	7	15
46	51	0	34	3	42	5
52	60	8	8	0	68	0
TOTAL	*%*	18.8	25.7	8.5	37.6	7.5
	$\bar{X}$	28.5	44.4	38.1	51.1	41.0

ITEM 4. ANOMALOUS

FRACTILE SCORE BANDS		"8"	1	2	3*	4
0	36	0	1	13	30	33
37	45	0	1	4	23	50
46	51	0	0	3	33	49
52	60	0	0	1	37	38
TOTAL	*%*	0.0	0.6	6.6	38.6	53.3
	$\bar{X}$	0.0	36.5	33.2	44.1	43.8

4. DISCUSSION

4.1. A comparison between tables 1 and 2

The first three columns provide the same information in both tables: Item, Subjects and Number of items correct. In the first part of this table we do not find relevant cues on the facility index. Most of the items are extremely easy as their p-value is very high and it is beyond the range Fulcher (1997) and Herrera (1999:209) suggested as reasonable. Only two items: 36 and 39 fall below .30, the lower boundary suggested. The information provided by the CTT, based on proportions, leads us to question the validity of most of the items if we intend to work with norm reference tests. We arrive at similar conclusions with the IRT. This time not from a proportion point of view but from a standardized metric perspective, where items are categorised according to their standard deviation. Thus, the examinees' behaviour on items 36 and 39 is so anomalous that they are not considered on the threshold scale. Most of the items are so easy that their level of difficulty falls far away from mean = *0* in a normal distribution. With this parameter items 4 and 18 are categorised as very difficult, being close to 3 SD, whereas items 8, 16, 19, 29 are categorised as extremely easy, since they are above 3 SD. None of these items should be used in a norm reference test, where discrimination is what really matters for this ability level.

It is assumed that the sample we are working with is not large enough to draw conclusions on the IRT two-parameter and three-parameter models. Nevertheless, the size of this pilot approach allows us to present some trends that are likely to be found when the sample fulfills the required parameters. Thus, it can be put forward that under the biserial correlation heading (table 1) useful information concerning the internal consistency of the item is found. Items 1, 4, 18, 36 and 39 are far from the >0.30 pointed out before. The slope column (table 2) shows a low discriminating power in almost half of the items, since the values observed should be between *1 and 2*. Under the asymptote column we learn that the guessing effect has been reasonable in all but item 18, which has reached 0.48, a disproportionate figure which could be explained on the basis of its difficulty.

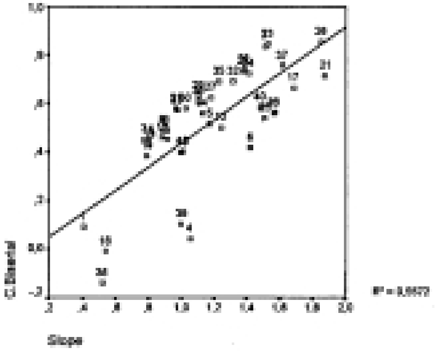

Fig. 2. Biserial and Slope correlations.

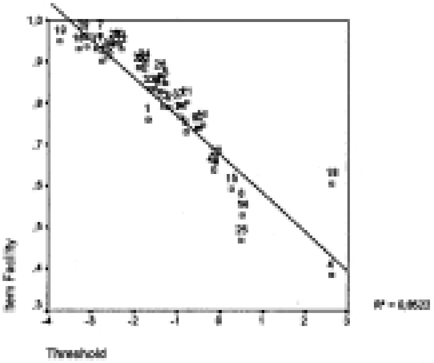

Fig. 3. Correlation: Item Facility and Treshold.

A better illustration of the examinees' performance can be observed in the above two figures. A high correlation is found between the biserial correlation and the slope (fig. 2) and between facility index and threshold (fig. 3). The negative correlation in the second figure must be interpreted as a result of relating facility of the CTT and difficulty of the IRT. Lack of consistency of some outlier items in relation to the rest of the test are shown in Fig. 2, where most of the items which have been previously labelled extremely easy or difficult or anomalous fall apart from the regression line. Fig. 3 also shows as outliers some of the items taken as such in Fig. 2. The determination coefficient (R^2) in both cases: threshold and item facility on the one hand and slope and biserial correlation on the other show a high correlation. These results demand for an in-depth analysis on some of these items from the fractiles perspective.

4.2. The contribution of fractiles to the assessment of the items

The most salient feature of this study is that none of the items included in the test belongs to the expected ideal model to which, from a theoretical point of view, every item should conform. The closest ones to the ideal paradigm fall within the so-called *dominant subcategory*, only 4 out of 60 items. The behaviour of distractors in the other *dominant categories* is quite distant from this ideal parameter, since there are quite a lot that are not considered at all, that is 43 out of 60. On the whole, that means that the low percentages in the distractors of the items identified as dominant were neither plausible, nor attractive alternatives since the correct option was chosen by an overwhelming majority of the candidates who took the test in a proportion higher than 0.70.

An analysis based on the data provided by fractiles will reveal some traits of the items that could shed some light on the candidates' behaviour.

a. The grammatical aspect considered is so easy that only motivating reasons could explain their inclusion in an upper intermediate test, where the test designer has resorted to texts and distractors in which we come across verbal structures or time prepositions beginners must deal with. (Items 16 and 29. Appendix)

b. The alternatives given to the correct option are, in a good deal of cases, out of place. Thus the probability of choosing the correct answer is enhanced. Examinees may not know the correct answer but on the grounds of partial knowledge they would not see any reason for choosing most of the distractors. It would be very easy for them to realise that the alternatives presented are not suitable to fill in the gap such as item 45 in which the distractors presented to a question, in which information about where an office is required, are out of context, or as happens in item 54 where the distractors presented belong to different grammatical categories: two modals and a preposition when an adverb is needed. (Item 54. Appendix).

c. Though Canale (1988) considers that it is very difficult to determine the boundaries of the different communicative competence levels, there is no doubt that the semantic solving problem proposed to examinees in item 7 again does not offer attractiveness or plausibility in the distractors for candidates that are supposed to be at an upper intermediate level. Sometimes the interest of distractors is based on misleading semantic interpretations that could arise from a term such as *book* (Appendix).

d. The combination of semantics and grammar leave no other alternative than the correct one (item 31.- Appendix). The other three options are out of place either because of the tense used in the distractor or because of the contradiction in presenting an aggrement and a negation at the same time.

Even though these commentaries refer to the *"exclusively dominant"* subclass they are applicable to the other three subclasses: *"dominant +1, dominant +2 and dominant +3"*, as also to the other categories described.

Finally, the items ascribed to the anomalous category warrant a specific commentary. It is likely that semantic misunderstanding between year and a specific date (item 4), misleading teaching on the use of *"in / at"* (item 18), a pragmatic clash between greeting and introducing people (item 36) or a matter of ignoring the issue of redundancy (item 39), could explain a higher frequency in the wrong option than in the correct option (Appendix).

On the whole, most of these situations could be explained on the grounds that too much importance was given to simple grammatical issues. Seven items were devoted to the use of common prepositions and a similar number of items asigned to the use of verb tenses and verb forms, all of which are too easy for upper intermediate candidates.

Distractors, in many cases are implausible and unattractive. Thus, candidates may have doubts as to which one is the correct option, but they do know what distractors are not to be considered as alternatives. Using partial knowledge they increase the probability of marking the right option and a 4 multiple choice test format may become a 3 or 2 multiple choice test format.

5. CONCLUSION

1. From the facility / difficulty index perspective, CTT shows two items to be outside normal parameters, while IRT not only shows items 36 and 39 to be anomalous but also that items 4 and 18 are extremely difficult. Items 8, 19, and 29 can be categorised as extremely easy, since they are above 3 Standard Deviations.

2. IRT enables us to determine the information level of each item on the different ability levels, but in our case the easiness of most of the items does not permit us to discriminate and provide information on the different levels of ability.
3. The schema we developed and worked with proved to be useful for the analysis of distractors.
4. The distractors used in the dominant category were neither plausible nor attractive alternatives.
5. Based on the sort of distractors offered it is likely that examinees, resorting to partial knowledge, marked the right option even when they did not know the correct answer.
6. Since the degree of difficulty is a matter not only of the item but also of the distractor, the test should be thoroughly revised and many distractors modified, mainly in the items commented on.
7. If the test was intended to measure a desired level for a personnel selection process, the test designer has not succeeded in providing even a single item where the correct option and appropriate distractors fulfill the expected criteria of a good item, at least for upper intermediate level candidates.
8. On the whole, these results invite us
   a) To carry out further research on the probability that a candidate achieves a correct response on the basis of his ability and the difficulty of the item in an IRT framework.
   b) To bear in mind that when piloting with a CTT framework, a test designer should, as far as possible, endeavour to achieve the ideal paradigm pointed out.

NOTES

[1] Item response theory has been and is being used in many large-scale testing programs trying to provide the information CTT was unable to.

[2] Fractiles divide scores into smaller groups of scores of approximately equal size. The median, quartiles, deciles or percentiles are fractiles where scores are divided into two, four, ten, or a hundred sets of scores respectively.

[3] Whereas Fulcher (1997) suggests a 40-60 range in the facility / difficulty index, Herrera (1999) spans this range to 30-70.

[4] X_c = R - [W/(m - 1)], where X_c is the corrected score, R is the number of right answers, W is the number of wrong answers, and m is the number of options in each of the multiple choice items in the test.

[5] ICADE students' are required a First English Certificate level to take this course and the "Escuela Oficial de Idiomas" students are on their fourth course, high above the First English Certificate.

[6] Rasch analyses on dichotomous items can be carried out with a recommended minimum of 20 such items (Wright and Stone, 1979) and approximately 100 subjects, though larger data sets will enable more precise estimates, whereas the two parameter model (discrimination) and three-parameter (guessing) models recommend larger set of data. (McNamara, 1996).

Honesto Herrera-Soler
Facultad CC.EE. y EE., UCM
Somosaguas, 28223 Madrid
Tel. 91 394 24 25
Fax 91 394 24 18
E-mail: hherrera@ccee.ucm.es

Rosario Martínez Arias
Facultad de Psicología, UCM
Somosaguas, 28223 Madrid
Tel. 91 394 30 58
E-mail: psmet01@sis.ucm.es

REFERENCES

Bliss, L.B. (1980). A test of Lord's assumption regarding examinee guessing behaviour on multiple choice tests using elementary school children. *Journal of Educational Measurement*, 17: 147:153.

Canale, M. (1988). The measurement of communicative competence. *Annual Review of Applied Linguistics*. 8:67-84.

Crocker, L. and J. Algina (1986). *Introduction to Classical and Modern Test Theory*. Iowa City: ACT.

Curtis, M.E. and R.F. Glaser (1983). Reading theory and the assessment of reading achievement. *Journal of Educational Measurement,* 20, 133-147

De Jong J.H. and A.L Stoyanova (1994). Theory building:sample size and data-model fit. Paper presented at the 12th annual Language Testing Research Colloquium, San Francisco, March, 1994.

De las Cuevas, Julián y Dalila Fasla (eds.) *Contribuciones al estudio de la Lingüística Aplicada*. La Rioja: Asociación Española de Lingüística Aplicada.

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

Griffin, P.E. (1985). The use of latent trait models in the calibration of tests of spoken language in large-scale selection-placement programs, in Lee Y.P., A.C.Y.Y. Fok, R.K. Hambleton, H. Swaminathan and H.J. Rogers (eds.). *New Directions in Language Testing: Papers Presented at the International Symposium on Language Testing*. Hong Kong, Oxford: Pergamon.

Hambleton, R.K., H. Swaminathan and H.J. Rogers (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Herrera-Soler, H. (1999). Lectura de los resultados en un test de elección múltiple, pp.207-215. En de las Cuevas, Julián y Dalila Fasla (eds.): 207-215.

Hoi, K.S. (1990). *Principles of Test Theories*. Hillsdale, N.J.: Lawrence Erlbaum Associates, Publishers.

Lazarsfeld P.F. and N. Henry (eds.). *Reading in Mathematical Social Science*. Chicago, IL: Science Research Associates, 89-107.

Lee Y.P., A.C.Y.Y. Fok, R.K. Hambleton, H. Swaminathan and H.J. Rogers (eds.) (1985). *Fundamentals of Item Response Theory*. Sage, Newbury Park, CA.

Lord, F.M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika.* 39: 247-264.

Lord, F.M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement,* 8: 147-151.

Ludlow, L.H. and O'Leary (1999). Scoring omitted and not-reached items: Practical Data Analisis Implications, *Educational and Psychological Measurement,* V.59, 4: 615-630.

McNamara, Tim. (1996). *Measuring Second Language Performance*. London: Longman.

Mislevy, R.J., and R.D. Bock (1989). *PC-Bilog 3: Item Analysis and Test Scoring with Binary Logistic Models*. Mooresville, IN: Scientific Software.

Rasch, G. (1966). An individualistic approach to an item analysis. In Lazarsfeld P.F. and N. Henry (eds.). *Reading in Mathematical Social Science*. Chicago, IL: Science Research Associates, 89-107.

Skehan, P. (1989). Language Testing. Part II. *Language Teaching* 22: 1-13.

Wilson, M..(1991). *Objective Measurement: Theory into Practice*. Norwood, NJ. Ablex.

Wilson, D.T., R. Wood, P.K. Downs and R.Gibbons. (1991). *TESTFACT: Test Scoring, Item Statistics, and Item Factor Analysis.* Chicago: Scientific Software, Inc.

Wright B.D. and M.H.Stone (1979), *Best Test Design*. MESA Press, Chicago, IL.

APPENDIX

1. There are two main areas in this project.

A. There could be two parts in this project.
B. There are at least two parts in this project.
C. There are only two parts in this project.
D. In this project one area is more important than the other.

4. By the following year we expect an increase of 5%.

A. Last year we had an increase of 5%.
B. In the years to come we expect an increase of 5%.
C. In the year after that one, we expect an increase of 5%.
D. In 2002 we expect an increase of 5%.

7. Her timetable is completely booked.

A. She has no free time at all.
B. She has a lot of free time.
C. She needs to go to the library.
D. Her timetable is written in a book.

8. I’m sure they will finish on time.

A. I’m positive. They will finish on time.
B. They can’t finish on time.
C. They may not finish on time.
D. They must finish on time.

11. If we don’t lower our interest rates, nobody will borrow money from us.

A. Unless you borrow money we won’t lower interest rates.
B. The lending of money depends on our lowering of interest rates.
C. If you lower interest rates, people won’t borrow money.
D. Unless we lower our interest rates, people will borrow money from us.

15. That’s perfect Tom, I’ll see you around 6:30pm.

A. I’d like to talk to you for a few more minutes now.
B. I have nothing more to say to you today.
C. I have no free time to see you later.
D. I don’t want to see you until our next meeting.

16. ______ you know how to send e-mails?

A. Doesn’t **B.** Aren’t **C.** Don’t **D.** Isn’t

18. Hello, I'm Tom Rivers; I work ______ the telecommunications department.

A. around **B.** at **C.** up **D.** in

19. The next linew of products will be the ______ of all.

A. best **B.** better **C.** good **D.** greater

25. Our company is ______ with Thompson Allied at the beginning of next year.

A. connecting **B.** connect **C.** merged **D.** merging

26. Thank you very much for ______ those plans to me so quickly.

A. getting **B.** got **C.** to get **D.** get

29. Don't be late for the meeting, it's _______ 8:00pm

A. at **B.** on **C.** in **D.** within

31. Good afternoon Mr. Kline, have you finished the project reports yet?

A. Yes, I received them yesterday.
B. *Yes, I just finished them an hour ago.
C. Yes, I'll finish them tomorrow.
D. Yes, my boss forgot to give them to me.

36. Hello, I'm Carl Turner, how do you do?

A. How are you, Carl?
B. I'm John, what's your name.
C. I'm fine thank you, my name is Peter.
D. I'm John Smith, I work in finance.

39. How often do you purchase overseas?

A. We sometimes buy overseas.
B. We have very good overseas suppliers.
C. We usually buy overseas.
D. We usually buy locally.

44. Can you take the subway to work?

A. Yes I take the subway.
B. No, the subway is faster.
C. No, it takes much too long.
D. No, I prefer public transport.

45. Pardon me, do you know where Mrs. Phillip’s office is?

A. Yes, she’ll be right with you.
B. Yes, down the hall, first door on the right.
C. No, she is out of the office right now.
D. No, I don’t know her address.

..... We are (52) _________ to purchase new software this year, for both our branches (53) in Chicago and Paris.

..... (54) __________ , if it is possible, we would like a representative to train (55) our employees in the application of this software.

Tom Rollins (C.E.O.) Allied Electronics

52. **A.** in the market **B.** look for **C.** about **D.** hope
54. **A.** also **B.** shall **C.** can **D.** until