FORMULACIÓN, DESARROLLO Y PERFECCIONAMIENTO DE PRUEBAS INICIALES PARA EL PRIMER CICLO DE ESTUDIOS EN TRADUCCIÓN DENTRO DE LA LICENCIATURA EN TRADUCCIÓN E INTERPRETACIÓN

[THE SPECIFICATION, DEVELOPMENT, AND REFINEMENT OF READING SKILL SUB-TESTS FOR INITIAL ENTRY TO TRANSLATION STUDIES WITHIN THE FIRST DEGREE PROGRAM IN TRANSLATION AND INTERPRETING.]

BRYAN ROBINSON

SUPERVISED BY DR J. N. D. MCLAREN

UNIVERSITY OF GRANADA
OCTOBER 1999

Editor: Editorial de la Universidad de Granada
Autor: Bryan Robinson
D.L.: Gr. 40- 2007
ISBN: 978-84-338-4218-3

ACKNOWLEDGEMENTS

This thesis has been long in writing – too long. It has been long in writing for many reasons, and has undoubtedly suffered for this delay. At the same time, both the author and many others — more or less directly involved — have also "suffered". At this late stage in the process it seems the very least I can do is to offer them my sincerest thanks:

- to Neil MacLaren, who agreed to direct this thesis when unbeknown to him it would take far longer than it should have;
- to Roberto Mayoral Asensio, who as Director of the EUTI and Dean of the FTI, gave me the opportunity to develop my thesis in the "messy" real-world
- to the many friends, colleagues and acquaintances in Granada, Cardiff, and elsewhere, who have participated in the research itself, or in its development, or who simply have offered much-needed encouragement at regular intervals;
- and, finally, to Josefina, Nicholas, and Lydia, whose time it has taken.

To all of them is due any credit that this work may hold. As for its faults, only I can take that responsibility.

B.J.R.

Granada, October 1999

Se defendió esta tesis doctoral, dirigida por el
Dr. D. J. N. D. McLaren, el día 4 de febrero de 2000,
ante un tribunal compuesto por:

Presidente	Dr. D. Enrique Alcaraz Varó
Secretario	Dr. D. Juan Santana Lario
Vocales:	Dr. D. Roberto Mayoral Asensio
	Dr. D. Christopher Waddington
	Dr. D. Juan Jesús Zaro Vera

La tesis obtuvo la calificación de *sobresaliente cum laude*.

RESUMEN

Contexto

En España nadie se opuso a la decisión de introducir una prueba específica de aptitud para la licenciatura en Traducción e Interpretación, a pesar de que previamente no se había definido el concepto de aptitud y tampoco se había demostrado empíricamente que existiera un u otro modelo de herramienta adecuado para medirlo. De hecho, la investigación publicada en este campo brillaba por su ausencia.

El propósito de nuestro trabajo

Dadas las carencias que encontramos, en este estudio correlacional abordamos el diseño, desarrollo y perfeccionamiento de unas subpruebas de destrezas lectoras en Lengua A Español (LA) y Lengua B Inglés (LB), que puedan formar parte de una batería de pruebas específicas de acceso, con la intención de que midan la aptitud para traducir del candidato. Nuestro primer propósito fue justificar el uso de una prueba específica para el acceso a los estudios en traducción y ratificar la hipótesis de que existe una relación directa entre las destrezas lectoras y la aptitud para traducir. El segundo propósito fue medir la relación entre nuestras subpruebas e instrumentos de evaluación global de la enseñanza secundaria y del conocimiento de LB. Estas relaciones figuran con frecuencia en estudios similares al nuestro (Spolsky 1995). El tercer propósito consistió en identificar cualquier combinación de destrezas de lectura y actividades de examen cuyo comportamiento se demostrara especialmente eficaz en el diseño de nuestro instrumento.

Método

Para ello, elegimos el modelo de diseño de investigación correlacional propuesto por J.B. Carroll (1981) y preparamos una serie de pruebas de destrezas lectoras basándonos en una especificación de prueba derivada de la taxonomía de destrezas comunicativas y lingüísticas de Munby (1978). Nuestro estudio se dividió en dos partes: una longitudinal y otra transversal. En la primera, calculamos el error estándar de medida (S_e) y la validez predictiva de nuestras pruebas para cumplir con la primera y la sexta de las condiciones que estipula el modelo de

Carroll. En la segunda, evaluamos la validez concurrente y criterial con respecto a una serie de medidas de aprovechamiento educacional y dominio general de LB.

Diseñamos las subpruebas según los criterios establecidos en la teoría clásica de examen (Bachman 1990). En este proceso utilizamos actividades y métodos de análisis del ítem conforme con los que se suelen encontrar en la bibliografía al uso (Weir 1983, 1990, 1993; Crocker y Algina 1986; Heaton 1988, 1999; Alderson et al 1995; Baxter 1997; entre otros).

Ensayamos nuestro diseño con una cohorte de alumnos de Diplomatura (n = 44), y la primera versión de las subpruebas de destrezas lectoras se administró (Cohorte A n > 440). Con los resultados de estas pruebas revisamos la especificación y administramos otras subpruebas (Cohorte B n > 350).

Resultados

Calculamos el error estándar de medida (S_e) de cada subprueba para evaluar la varianza de la puntuación verdadera en la actuación de la prueba. Ninguna de las subpruebas alcanzó el nivel bajo que habíamos fijado como meta ($S_e < 1$), y la actuación de todos hubiera mejorado al incluir mayor número de ítems (entre +6 y +14).

Las correlaciones bajas y positivas, de significación estadística, que calculamos al comparar nuestras subpruebas en LB con otros instrumentos de medición que habíamos fijado, demostraron la validez predictiva de nuestras subpruebas. Los instrumentos con los que las comparamos fueron las notas finales de las asignaturas de Traducción General A-B y B-A (Cohorte A, r = 0.45; 0.25, respectivamente). Los datos no significativos generados por la Cohorte B tienden a confirmar los resultados, a pesar de que éstos se vieron afectados por otras variables que no pudimos controlar. Aunque ninguna de las subpruebas consiguió los niveles que habíamos fijado, la mayoría de las correlaciones fueron del rango entre 0.2 y 0.6, el que típicamente se encuentra en este tipo de estudio (Spolsky 1995).

En nuestro análisis de las subpruebas de LA, solo pudimos obtener datos para la Cohorte A, y éstos no confirmaron nuestra hipótesis. Las correlaciones (LA:A-B, r = 0.13; LA:B-A, r = -0.09) parecen indicar diferencias importantes entre los criterios que se aplican a la evaluación de traducción hacia las dos lenguas. Pensamos que la producción escrita pudiera ser un instrumento óptimo para medir la aptitud para traducir hacia LA, siempre y cuando se pueda asegurar la objetividad adecuada en las subpruebas pertinentes. Nos proponemos investigar en este campo en el futuro.

Medimos la validez concurrente de nuestras subpruebas, comparándolas con los datos obtenidos de dos instrumentos de evaluación de rendimiento académico de los candidatos: la nota de Selectividad (Univ) y la media global de COU (Ssch). Nuestro propósito fue el de confirmar o rechazar la necesidad de una prueba específica de acceso. Las correlaciones generadas fueron muy bajas, positivas, y cero, para la relación entre Univ y los exámenes de traducción tests (Cohorte A: Univ:A-B r = 0.22; Univ:B-A r = 0.11; Cohorte B: Univ:A-B r = 0.06; Univ-B-A r = 0.00). A continuación nos sorprendió que tres de las cuatro correlaciones que calculamos, entre la media de COU y los exámenes de traducción, fueran negativas, y que en el caso de la Cohorte B la correlación entre estas dos notas (Univ:Ssch) no pasara el 0.46. Sugerimos que la naturaleza de estas correlaciones, entre estas dos notas, indica que una investigación de los criterios que se aplican en la enseñanza secundaria y en las pruebas de acceso a la universidad aportaría resultados beneficiosos.

En la evaluación de la validez criterial, encontramos que la prueba publicada, la *Oxford Placement Test* (Allan 1990) resultó más eficaz como instrumento para medir la actuación de los candidatos que nuestras subpruebas. Ambas cohortes (A, r = 0.35; B, r = 0.45) generaron datos que indican un solapamiento de destrezas entre el examen de inicio en Inglés general y nuestras subpruebas, de modo que la validez predictiva resultó mejor: Cohorte A : Transl. A-B r = 0.67; Transl. B-A r = 0.17; Cohorte B Transl. A-B r = 0.25; Transl. B-A r = 0.29). Evidentemente, no es factible utilizar una prueba que se comercializa como prueba de acceso, no obstante, los valores y la naturaleza de estas correlaciones nos llevan a unas conclusiones significativas. De entrada, como descubrió Clapham (1975), conocimientos lingüísticos generales y destrezas lingüísticas parecen indicar aptitud, mucho mejor que pruebas más restringidas, como las nuestras. En segundo lugar, la naturaleza de las destrezas que nosotros evaluamos en los exámenes de traducción general parecen ser más complejas. Una amplia batería de subpruebas, que incluya aspectos de conocimientos gramaticales, expresión escrita, actividades de corrección y precisión, además de las destrezas lectoras que hemos evaluado, parece ser la más adecuada.

La decisión de elegir las destrezas comunicativas descritas por Munby parece haberse justificado en el grado de éxito de muchos de los ítems, a pesar de que el análisis posterior indica a menudo que las destrezas en las que nos hemos fijado no han resultado ser las únicas que se utilizan en el proceso de responder al estos ítems.

Comentario

Pocos aspectos de nuestra investigación dan resultados definitivos y, en su conjunto esto es común dentro de lo complejo que resulta participar en la industria de los exámenes a gran escala, tal y como las pruebas específicas de acceso implican. Nuestro estudio generó correlaciones similares a las que se encuentran en la bibliografía, aunque a nosotros nos han decepcionado. No obstante, como anteriormente comentamos, este estudio abre las puertas de la investigación a muchos campos nuevos. Esperamos que nuestra labor estimule a otros, en el campo de los estudios de traducción, a *darse el gusto* de preparar trabajos empíricos similares.

ABSTRACT

Background

While the decision to use a specific entry test for first degree studies in Translation and Interpreting in Spain was uncontested, neither the concept of aptitude for translation had been defined, nor a satisfactory test instrument design had been empirically proven. Indeed, published research in the area was conspicuous by its absence.

Aims

In this correlational study, we set out to design, develop, and refine reading skills sub-tests in Language A Spanish (LA) and Language B English (LB) to form part of a battery of entry tests intended to measure aptitude for translation. We aimed to justify the use of a specific entry test for Translation studies, and to test the hypothesis that a direct relationship exists between reading skills and aptitude for translation. We also intended to measure the relationship between our sub-tests and measures of secondary school achievement and LB proficiency, typical of those used in similar research projects (Spolsky 1995). A further aim in our research was to identify any combinations of reading skills and test task types, which might prove particularly efficient in our test instrument design process.

Method

We adopted the correlational research design model proposed by J.B. Carroll (1981) and prepared a series of reading sub-tests based on a specification derived from Munby's taxonomy of communicative language skills (1978). We divided our study into two parts: one longitudinal, and the other cross-sectional. In the former, we assessed the Standard error of measurement (S_e) and the predictive validity of our tests in order to fulfil conditions one and six of Carroll's model. In the latter, we assessed concurrent and criterion validity against a set of measures of educational achievement and general LB proficiency.

We designed our tests according to criteria established in Classical Test Theory (Bachman 1990). In the process, we used test task types and methods of item analysis in line with those

widely described in the literature (Weir 1983, 1990, 1993; Crocker and Algina 1986; Heaton 1988, 1999; Alderson et al 1995; Baxter 1997; among others).

Our test specification was trialled on a cohort of diploma course students (n = 44), and the first full version of the reading sub-tests was administered to Cohort A (n >440). Based on these results we revised the specification, and administered further sub-tests to Cohort B (n >350).

Results

We calculated the Standard error of measurement (S_e) of each sub-test as a means of estimating the degree of true score variance in test performance. None of the individual sub-tests achieved the low S_e levels we aimed for (< 1), and all of them would have been improved by the inclusion of more items (range +6 to +14).

The predictive validity of our LB sub-tests was proven by the statistically significant, low positive correlations obtained between these and the target performance measures. The instruments used for this comparison were the course final examinations in General Translation A-B and B-A (Cohort A, r = 0.45; 0.25, respectively). In part, the non-significant data for Cohort B supported these results, although they were influenced by variables in the test administration, which were beyond our control. Although none of the tests achieved the target values we had established, most of the correlations generated were within the 0.2 to 0.6 range, which is typical of similar studies (Spolsky 1995).

For our analysis of the LA sub-tests, we could only obtained data for Cohort A, and these did not support our hypothesis. The correlations (LA:A-B, r = 0.13; LA:B-A, r = -0.09) would seem to indicate important differences in the criteria used to assess translation into each of the languages. We now believe that written production may well prove a better instrument by which to measure LA aptitude for translations, always provided that adequate objectivity can be achieved. We propose further research in this area.

We measured the concurrent validity of our sub-tests against two scores representing candidates' academic achievement: the University entrance test score (Univ), and the Secondary school average (Ssch). We did this to establish whether or not a separate test was justified, and generated very low positive and zero correlations between Univ and the translation tests (Cohort A: Univ:A-B r = 0.22; Univ:B-A r = 0.11; Cohort B: Univ:A-

B $r = 0.06$; Univ-B-A $r = 0.00$). We were surprised, then, to find that three of the four correlations between Ssch and the translation examinations were negative, and in the case of Cohort B the correlation between these two instruments, (Univ:Ssch), was only 0.46. These data lead us to conclude that a specific test of aptitude is justified. Furthermore, we would suggest that the nature of the correlations between these measures suggests research into the differing criteria applied in secondary schooling and university entrance tests would bear fruit.

In assessing criterion validity, we found that the commercially prepared Oxford Placement Test (Allan 1990) was a more successful measure of candidate performance than our sub-tests had been. Cohorts A ($r = 0.35$) and B ($r = 0.45$) both provided us with data indicating an overlap of skills between the General English placement test and our own sub-tests, and their predictive validity was better: Cohort A : Transl. A-B $r = 0.67$; Transl. B-A $r = 0.17$; Cohort B Transl. A-B $r = 0.25$; Transl. B-A $r = 0.29$). Although it is clearly impractical for us to use a test that is openly available on the market as an entry test, the strengths and nature of these correlations lead us to significant conclusions. Firstly, as Clapham (1975) found, general language knowledge and skills would appear to be a better indicator of aptitude than a more restricted test such as ours. Secondly, the nature of the skills that we are assessing in the General Translation examinations would appear to be more complex. A wide ranging battery of sub-tests, that include aspects of grammar knowledge, written production, accuracy and precision tasks, in addition to the reading skills we have tested, would seem to be more appropriate.

Our decision to adopt Munby's skills as the basis for test design would appear to be vindicated by the levels of success of many of the items. We affirm this, despite the fact that subsequent analysis often indicates that the target skills are not the only ones candidates may have used in the process of responding to the items.

Commentary

Our research has proved inconclusive in a number of respects, and has demonstrated many of the complexities involved in taking part in the language testing industry on a scale as large as that which the specific entry test entails. The levels of correlation our study has provided are in line with those presented in the literature, although they remain disappointing for us. However, as we have mentioned, this study has opened a number of areas for further research. We would hope that it has encouraged others in the field of Translation studies to indulge themselves in the preparation of similar empirical studies.

RESUMEN DE CONTENIDOS

AGRADECIMIENTO iii
RESUMEN v
RESUMEN DE CONTENIDOS xiii

CHAPTER 1 INTRODUCTION 1
CHAPTER 2 ANNOTATED DEFINITIONS 13

Part I Background

CHAPTER 3 CONTEXT 27
CHAPTER 4 "APTITUDE IS NOT EASY TO DEFINE" 55
CHAPTER 5 THE TRANSLATOR "AS EXPERT READER" 81
CHAPTER 6 THE STARTING POINT FOR OUR TEST DESIGN: MUNBY, AND THE DESCRIPTION OF READING SKILLS 101
CHAPTER 7 CLASSICAL TEST THEORY 111

Part II Research design and methodological considerations

CHAPTER 8 APPROACHES TO RESEARCH DESIGN: CORRELATIONAL RESEARCH IN THE VALIDATION OF APTITUDE TESTS 135
CHAPTER 9 INSTRUMENT DESIGN 163

Part III Results

CHAPTER 10 TEST DEVELOPMENT AND REFINEMENT 177
CHAPTER 11 SUB-TEST ES01.1 185
CHAPTER 12 SUB-TEST IN01.1 199
CHAPTER 13 SUB-TEST ES01.2 213
CHAPTER 14 SUB-TEST ES02 231
CHAPTER 15 SUB-TEST IN02 255
CHAPTER 16 SUB-TEST ES04 281
CHAPTER 17 SUB-TEST IN03 311
CHAPTER 18 MEETING CRITERIA: THE STRENGTHS AND NATURE OF RELATIONSHIPS BETWEEN MEASURES 339

Part IV Discussion

CHAPTER 19 CONCLUSIONES 369
BIBLIOGRAPHY 383
ERRATA 398
APPENDICES 399

BRIEF CONTENTS

ACKNOWLEDGEMENTS iii
ABSTRACT ix
BRIEF CONTENTS xv
DETAILED CONTENTS xvii

CHAPTER 1 INTRODUCTION 1
CHAPTER 2 ANNOTATED DEFINITIONS 13

Part I Background

CHAPTER 3 CONTEXT 27
CHAPTER 4 "APTITUDE IS NOT EASY TO DEFINE" 55
CHAPTER 5 THE TRANSLATOR "AS EXPERT READER" 81
CHAPTER 6 THE STARTING POINT FOR OUR TEST DESIGN: MUNBY, AND THE DESCRIPTION OF READING SKILLS 101
CHAPTER 7 CLASSICAL TEST THEORY 111

Part II Research design and methodological considerations

CHAPTER 8 APPROACHES TO RESEARCH DESIGN: CORRELATIONAL RESEARCH IN THE VALIDATION OF APTITUDE TESTS 135
CHAPTER 9 INSTRUMENT DESIGN 163

Part III Results

CHAPTER 10 TEST DEVELOPMENT AND REFINEMENT 177
CHAPTER 11 SUB-TEST ES01.1 185
CHAPTER 12 SUB-TEST IN01.1 199
CHAPTER 13 SUB-TEST ES01.2 213
CHAPTER 14 SUB-TEST ES02 231
CHAPTER 15 SUB-TEST IN02 255
CHAPTER 16 SUB-TEST ES04 281
CHAPTER 17 SUB-TEST IN03 311
CHAPTER 18 MEETING CRITERIA: THE STRENGTHS AND NATURE OF RELATIONSHIPS BETWEEN MEASURES 339

Part IV Discussion

CHAPTER 19 CONCLUSIONS 376
BIBLIOGRAPHY 383
ERRATA 398
APPENDICES 399

DETAILED CONTENTS

ACKNOWLEDGEMENTS iii

RESUMEN v

ABSTRACT ix

RESUMEN DE CONTENIDOS xiii

BRIEF CONTENTS xv

DETAILED CONTENTS xvii

CHAPTER 1 INTRODUCTION 1

1.1 REAL-WORLD TOPICS ARE INHERENTLY "MESSY" 2
1.2 "SOCIETY'S GATEKEEPERS": THE ROLE OF AN ENTRY TEST 3
1.3 A BRIEF OUTLINE OF OUR RESEARCH 4
1.4 OVERALL RESEARCH QUESTION 8
1.5 SUMMARY 10

CHAPTER 2 ANNOTATED DEFINITIONS 13

Part I Background

CHAPTER 3 CONTEXT 27

3.1 AN OVERVIEW OF THE CONTEXT 28
3.2 A NEW OPPORTUNITY 29
3.3 PUBLIC DEBATE OVER PROGRAMS 29
3.4 THE DIPLOMA EXPERIENCE: AT THE GRASS ROOTS LEVEL 30
3.5 PREDICTED CANDIDATE NUMBERS 39
3.6 ADMISSION TO TRANSLATION STUDIES AROUND THE WORLD 40
3.7 CONCLUSIONS 53

CHAPTER 4 "APTITUDE IS NOT EASY TO DEFINE" 55

4.1 A HISTORY OF APTITUDE TESTING 57
4.2 AN OVERVIEW OF FLL APTITUDE 61
4.3 JOHN B. CARROLL: APTITUDE TESTING AND RESEARCH 64
4.4 THE MLAT, STILL IN USE IN THE 90S 75
4.5 "HORSES FOR COURSES": AN EXAMPLE OF AN APTITUDE TEST DESIGNED FOR A HIGHLY SPECIFIC POPULATION 77
4.6 APTITUDE FOR TRANSLATION 79
4.7 CONCLUSIONS 79

CHAPTER 5 THE TRANSLATOR "AS EXPERT READER" 81

5.1 TRANSLATION AND LITERATURE 83
5.2 THE READING PROCESS 86
5.3 FACTORS THAT INFLUENCE THE TRANSLATION PROCESS 87
5.4 QUALITATIVE RESEARCH INTO THE TRANSLATION PROCESS 91
5.5 A PROCESS-ORIENTED APPROACH TO TRANSLATOR TRAINING 93
5.6 CONCLUSIONS 99

CHAPTER 6 THE STARTING POINT FOR OUR TEST DESIGN: MUNBY, AND THE DESCRIPTION OF READING SKILLS 101

6.1 THE DESCRIPTION OF READING SKILLS 108
6.2 CONCLUSIONS 110

CHAPTER 7 CLASSICAL TEST THEORY 111

7.1 MAJOR TRENDS IN TESTING 112
7.2 CLASSICAL TEST THEORY 115
7.3 FACILITY VALUE 130
7.4 DISCRIMINATION 130
7.5 AN ALTERNATIVE TO CTT?: ITEM RESPONSE THEORY 130
7.6 CONCLUSIONS 131

Part II Research design and methodological considerations

CHAPTER 8 APPROACHES TO RESEARCH DESIGN: CORRELATIONAL RESEARCH IN THE VALIDATION OF APTITUDE TESTS 135

8.1 RESEARCH DESIGN 136
8.2 CARROLL'S SIX CONDITIONS TO TEST THE VALIDITY OF A MEASURE OF APTITUDE 139
8.3 TEST DESIGN 141
8.4 STEPS IN THE TEST-WRITING PROCESS 144
8.5 SAMPLING 157
8.6 VALIDATION AND TEST ANALYSIS 160
8.7 CONCLUSIONS 161

CHAPTER 9 INSTRUMENT DESIGN 163

9.1 MATCHING EXERCISES 164
9.2 DICHOTOMOUS ITEMS 165
9.3 MULTIPLE CHOICE QUESTIONS 166
9.4 ERROR IDENTIFICATION AND ERROR CORRECTION 168
9.5 SUMMARY WRITING 170
9.6 UNIVERSIDAD PONTIFICIA DE COMILLAS 172
9.7 CONCLUSIONS 172

Part III Results

CHAPTER 10 TEST DEVELOPMENT AND REFINEMENT 177

10.1 THE TEST DEVELOPMENT AND REFINEMENT PROCESS 178
10.2 A NOTE ABOUT PRESENTATION 180

CHAPTER 11 SUB-TEST ES01.1 185

11.1 TEXT 186
11.2 TASK TYPES 186
11.3 CONCLUSIONS 196

CHAPTER 12 SUB-TEST IN01.1 199

12.1 TEXT 200
12.2 TASK TYPES 202
12.3 RELIABILITY 210
12.4 DESCRIPTIVE STATISTICS 211
12.5 CONCLUSIONS 211

CHAPTER 13 SUB-TEST ES01.2 213

13.1 TEXT 214
13.2 TASK TYPES 216
13.3 DESCRIPTIVE STATISTICS 225
13.4 RELIABILITY 227
13.5 CONCLUSIONS 227

CHAPTER 14 SUB-TEST ES02 231

14.1 TEXT 232
14.2 TASK TYPES 235
14.3 DESCRIPTIVE STATISTICS 248
14.4 RELIABILITY 248
14.5 CONCLUSIONS 250

CHAPTER 15 SUB-TEST IN02 255

15.1 TEXT 256
15.2 TASK TYPES 256
15.3 DESCRIPTIVE STATISTICS 274
15.4 RELIABILITY 274
15.5 CONCLUSIONS 277

CHAPTER 16 SUB-TEST ES04 281

16.1 TEXT 283
16.2 TASK TYPES 285
16.3 DESCRIPTIVE STATISTICS 305
16.4 RELIABILITY 307
16.5 CONCLUSIONS 309

CHAPTER 17 SUB-TEST IN03 311

17.1 TEXT 315
17.2 TASK TYPES 315
17.3 DESCRIPTIVE STATISTICS 333
17.4 RELIABILITY 333
17.5 CONCLUSIONS 337

CHAPTER 18 MEETING CRITERIA THE STRENGTHS AND NATURE OF RELATIONSHIPS BETWEEN MEASURES 339

18.1 ARE THE SUB-TESTS WE DESIGN AND USE IN THIS STUDY VALID? 341
18.2 ARE THE SUB-TESTS WE DESIGN AND USE IN THIS STUDY RELIABLE? 358
18.3 DOES THE DATA GENERATED IN OUR RESEARCH LEAD TO ANY OTHER SIGNIFICANT CONCLUSIONS? 359
18.4 ITEM ANALYSIS 362

Part IV Discussion

CHAPTER 19 CONCLUSIONES 369

19.1 ¿SE JUSTIFICA EL USO DE UNA PRUEBA ESPECÍFICA DE APTITUD? 370
19.2 ¿LENGUA A O LENGUA B? ¿ACTIVA OR PASIVA? 370
19.3 ¿ANÁLISIS O SÍNTESIS? 371
19.4 EL VALOR Y LA NATURALEZA DE LOS COEFICIENTES DE CORRELACIÓN 372
19.5 ¿SE PUEDE MEDIR APTITUD PARA TRADUCIR A TRAVÉS DE LAS DESTREZAS LECTORAS? 372
19.6 CUÁLES SON LAS MODIFICACIONES QUE PODEMOS INTRODUCIR EN LAS SUBPRUEBAS PARA MEJORARLAS? 374
19.7 ¿QUÉ MODIFICACIONES PROPONEMOS PARA LA BATERÍA DE PRUEBAS? 374
19.8 ¿CUÁLES SON LOS CAMPOS QUE SE PUEDEN INVESTIGAR EN EL FUTURO? 374

CHAPTER 19 CONCLUSIONS 376

19.1 IS A SPECIFIC ENTRY TEST JUSTIFIED? 377
19.2 LANGUAGE A OR LANGUAGE B? ACTIVE OR PASSIVE? 377
19.3 ANALYTIC OR SYNTHETIC? 378
19.4 THE STRENGTH AND NATURE OF CORRELATIONCOEFFICIENTS 379
19.5 CAN APTITUDE FOR TRANSLATION BE MEASURED BY READING SKILLS? 379
19.6 WHAT MODIFICATIONS CAN WE INTRODUCE TO IMPROVE THE SUB-TESTS? 380
19.7 WHAT MODIFICATIONS DO WE PROPOSE FOR THE BATTERY OF TESTS? 380
19.8 WHAT FURTHER RESEARCH IS NEEDED? 381

BIBLIOGRAPHY **383**

ERRATA **398**

APPENDICES **399**

SUB-TEST IN01.1 401
TEST SPECIFICATION COHORT A 411
SUB-TEST ES01.2 415
SUB-TEST ES02 441
SUB-TEST IN02 459
TEST SPECIFICATION COHORT B 478
SUB-TEST ES04 483
SUB-TEST IN03 503

1 Introduction

1 INTRODUCTION

1.1 Real-world topics are inherently "messy"
1.2 "Society's gatekeepers": the role of an entry test
1.3 A brief outline of our research
1.4 Overall research question
1.4.1 *Hypotheses*
1.4.2 *The null hypothesis*
1.5 Summary

1.1 Real-world topics are inherently "messy"

This research project arose as the answer to a real-world challenge. The creation of a first-degree program in Translation and Interpreting, to replace the diploma level studies taught previously, meant that those centers that chose to could employ an entry test of APTITUDE. In this study, we define such a test as

> *A test designed to predict or measure a candidate's potential for success within a particular area of learning, e.g. in learning a foreign language, or on a specific course of study. (ALTE 1998:135)*

This definition, and all the other terms that we present in small capitals, appears in **Chapter 2** "Annotated definitions".

Our decision to make the entry test the focus of a doctoral thesis clearly flew in the face of the advice of experienced writers and researchers (Phillips & Pugh 1994:50) who firmly counsel against choosing problem-solving research. They rightly point out that real-world topics are inherently "messy", as they can seldom be limited to the confines of a single discipline. However, inter- or multi-disciplinarity is one of the most striking characteristics of the field of Translation studies (Snell-Hornby 1994; Bowker et al 1998). Once we accepted this, the range of research questions arising from this challenge seemed, at the very least, coherent with the discipline itself.

Translation studies is as old as the "oldest professions" (Snell-Hornby 1994a; Beeby Lonsdale 1996) and yet it is only in the last decade that the formal discipline has been recognized as such. In Spain, legal status was achieved when the *área de concocimiento* in Applied linguistics for translation and interpreting was established in 1993. To this extent, Translation studies derives its multidisciplinary character, we would suggest, from the multidisciplinary backgrounds of those teachers and researchers who have made it their academic "home".

Coming from the world of the professional translator, of foreign or second language teaching, linguistics - whether applied or theoretical - literary studies, or wherever, those working in the field of translation have inevitably brought with them the baggage of their training and knowledge. They have applied these varied disciplines to research questions arising within Translation studies, and have offered many revealing perspectives of the area. Increasingly the field has been considered inter-disciplinary, with researchers and theorists describing a holistic approach, wherein the professional translator is viewed as a specialist both in subject matter and in written production (Snell-Hornby 1994).

Whether the situation is a weakness or a strength is unclear, and recent work on this very question leaves the matter open (Bowker et al 1998). It is not surprising, however, that our study should have drawn together a range of academic fields. In order to carry out this project we built on experience, and we collated, classified, and systematically quantified data in our attempt to provide insights into a particularly thorny question which involves both academic and practical, real-world issues. In our study we call on research in Second and Foreign Language Learning, Applied Linguistics, Test theory, Statistics, and Translation Studies. In addition, we draw on the wide range of Information Technology skills we required in order to process data produced by over 700 candidates, each of whom took both of the reading skills sub-tests on which we base our study.

We acknowledge that the very broad scope of this work is one of its limiting factors. By attempting to cover a number of related but distinct areas, we have not been able to attain the depth of study that a more narrowly defined project would have permitted. We recognize this as a significant failing.

1.2 "Society's gatekeepers": the role of an entry test

In September 1992, the Spanish central government authorized the use of a test of aptitude for translation as a means of selecting candidates for entry to the newly created first degree in Translation and Interpreting. This action closed a lengthy debate during the course of which an entry test had been agreed upon and then rejected in the belief that it breached the Spanish constitutional right to non-discriminatory access to education. The degree of specialization involved Translation studies — similar to that in Fine Arts and Physical Education — was the justification for the test.

This debate leads us to the first hypothesis on which we base our study, and to the corresponding null hypothesis:

H Translation studies involves a degree of specialization such that a specific entry test of aptitude is justified

H_0 Translation studies does not involve a degree of specialization so great as to warrant the use of a specific entry test of aptitude

If we are to test, which means setting ourselves up as another of "society's gatekeepers", then we must be sure that there is an academic purpose to our test. Moreover, we have an ethical obligation to demonstrate that we are not simply introducing a bureaucratic, and economic hurdle into an already complex University entrance procedure.

1.3 A brief outline of o ur research

In **Part I** of our study, we report on the background against which we developed the theoretical and practical aspects of our work. In five chapters, we move from the historical context in which the need for an entry test arose to our search for an appropriate definition of aptitude and for an adequate research design on which to model our study. We found both of these in the extensive writings of J.B. Carroll (1974, 1978, 1981, 1993; Carroll, J.B. and Sapon 1959). We continued our research by touching on areas of translation theory and translation pedagogy, which led us to the taxonomy of communicative language skills drawn up by Munby (1978). From this list, we selected the core elements of our test instrument specification. Finally, we summarize the essential aspects of Classical Test Theory and highlight those we have chosen as the empirical frame within which to carry out our quantitative, correlational study.

We begin then, with a look at the context of our research, in **Chapter 3**. The chapter is in two parts. The first deals with the quantification of previously uncollated data derived from student performance at entry to, and after the first year of studies in the Diploma course subject *Traducción I*. Here we describe the scandalously high failure rates that existed in this subject, and look at two attempts at damage limitation, which produced mixed results. We study data gathered from the Diploma course students that later serve as benchmark standards against which we evaluate our research.

In the second half of the chapter we move outside of Spain and seek published research from around the world. Here we find the work of Bossé-Andrieu (1981) and her colleagues Campagna & Dionne (1981). Their work throws up a number of issues that have a bearing on the Spanish context, and which will feature in the design of our study.

Experience, and the criticisms of professional translators outside of the academic world, were the two principal motives cited in Bossé-Andrieu's correlational study of the process of admission to translation schools in the francophone world (1981). The author depicted a lack of empirical analysis in the application of a range of aptitude tests in Belgium, Canada, France, and Switzerland. In her research, Bossé-Andrieu found that any definition of "aptitude for translation" was implicit in the instruments used rather than explicitly stated in a test specification. In the countries she surveyed, aptitude was equated with intellectual capacity and written production in LANGUAGE A (LA) — the language of habitual use: "de bien penser y de bien écrire". In contrast, in our research we find that aptitude, in Granada at least, is more often equated with proficiency in LANGUAGE B (LB), the first foreign language.

As we have said, the real-world experience on which we base this study is that of the years of translation teaching and learning at the University of Granada, during which LB proficiency was evidently of paramount importance. Once we had quantified this factor, we added it to the data, and to conclusions presented by Bossé-Andrieu (1981).

We conclude Chapter 3 with a number of unanswered questions as regards essential elements of entry testing specifications. The first has to do with the primacy of the Language A or the Language B in the entry procedure. The second centers on the testing of the active skills associated with written production or the so-called "passive" skills associated with reading. Finally, we encounter the very real dilemma over the choice of an ANALYTIC or a SYNTHETIC approach to the test instrument design. We see that the emphasis in Granada has been on the combination of LB, active skills, and a synthetic test instrument. In contrast, it would seem that in the francophone world the balance tends to towards a combination of LA, the active skills, and an analytic approach.

Our major criticism of the synthetic approach as it was applied in Granada before 1992 lies in the unscientific nature of its use. We believe that this came from an over-zealous interpretation of the much cherished *libertad de cátedra*. We hold that Classical Test Theory's concept of PRACTICALITY, and the challenge represented by testing a potential maximum of 700 candidates, obliges us to adopt an analytic approach to the entry test. We believe that, at the very least, this is the most politically correct decision. By making every effort to achieve maximum objectivity and reliability in the design and use of our test instruments, we hope to avoid some of the inherent difficulties of the task before us.

In **Chapter 4**, we move beyond the confines of our historical and disciplinary context, and search in the literature on general linguistic aptitude. We do so because, although the essential concept described in our research is aptitude for translation, this has been examined in little detail in the literature reviewed, and from very diverse positions; (Bossé-Andrieu 1981; Campagna and Dionne 1981; Atsuko, 1988). Most teachers of translation appear to share an implicit definition of aptitude, but in order to select an appropriate research model we read within the fields of Foreign language learning (FLL) and Second language learning (SLL) research. Here, we found that aptitude and its measurement have exercised the minds of researchers for many years. J.B. Carroll (1974, 1978, 1981, and 1993), Carroll, J.B. and Sapon (1959) Gardner and Lambert (1965, 1972), Ellis (1986), Skehan (1986) and Ehrman (1995) among many others, have all contributed.

In SLL and FLL, aptitude is one of the five variables believed to influence the rate of language learning (Stern 1983, Ellis 1986). Since the early years of this century, aptitude has been the objective of many tests which, almost without exception, have been aimed at keeping "prospective failures out of classes" (Spolsky 1995:117). The means of achieving this have been by using measures of aptitude as entry or placement tests. As such they, and we in our turn, take on the role of "society's gatekeepers", and this carries a significant ethical burden.

The widespread use of BATTERIES of SUB-TESTS has been among the factors that we have found these tests have in common. In order comprehensively to evaluate a range of elements considered components of the target aptitude, test designers reveal a widespread belief that aptitude is in essence a divisible conglomerate of skills or abilities. They test these through discrete exercises, each of which focuses on a specific area. The sum score of these tests then provides the overall test score. Another common element is the correlation of scores produced by criterion validation of these tests through comparison with indicators of IQ, or general intelligence often measured through secondary school grades or University entrance examinations.

The work of J.B. Carroll draws together all of the principal characteristics of aptitude research. From Carroll, we derived the research design model that this study is based on. In Chapter 4, we discuss a selection of his major research findings, and analyze the manner in which he has viewed different aspects of aptitude, especially in order to distinguish between general intelligence and other aptitudes. We also look at the structure and use of a number of aptitude tests, and find that the strength and nature of the correlations researchers have produced tend to range from 0.20 to 0.60. Furthermore, we find that in work that is more recent, Ehrman (1992) accepts the lower end of this range as being a satisfactory result given the highly homogeneous nature of the cohort on which her study was based.

Having reached the conclusion that the concept of aptitude as a divisible entity is the more widely accepted, although empirically unproven standpoint from which to begin our instrument design, we move in **Chapter 5** into the field of translation studies. Here we follow the search for a description of aptitude into two camps. We begin by looking at the concept of a unitary aptitude, propagated among many literary translators. Then we read into Translation theory (Bell 1991; Reiss 1992; Lörscher 1986, 1991, 1992, where we find the opposite view expounded in research into the translation process. From our readings in this field, and in translation pedagogy (Beeby Lonsdale 1996; Gile 1995), we conclude that the use of a taxonomy of communicative language skills is a satisfactory starting point for the development of a test specification. Furthermore, we found that within the pre-translation phase, the ability to read as an "expert reader" is crucial. Therefore, our proposed battery of tests would include discrete sub-tests in reading skills, among others.

This decision leads us, in **Chapter 6**, to describe the background against which Munby (1978) produced his taxonomy of communicative language skills, and to present those skills that we consider the essential base from which to write our test specification. Our test instrument model applies to the writing of sub-tests in Language A Spanish, and in the three Languages B: English, French and German. However, in this study we only consider the results of the Spanish and English reading skills sub-tests.

Having arrived at this point, we use **Chapter 7** to present an overview of the major concepts in Classical Test Theory (CTT), which we use as the empirical frame within which to apply our research design. CTT concepts, and the mechanisms of item analysis on which CTT is based are defined and described in this chapter, and we outline the manner in which we hope to apply them.

In SL research, or Social Science research, the quantitative versus qualitative debate over approaches to research design goes on. In this study, we use a correlational research design from the quantitative end of the continuum (Larson-Freeman and Long 1991) and adopt both descriptive and quasi-experimental methods of data analysis. The most important of the quantitative aspects of this study are found in the data collection, and in the application of CT Theory (Alderson et al. 1995; Bachman 1991). We analyze and correlate raw scores attained by candidates in our LA and LB reading skills sub-tests, with measures relating to Secondary school performance, the University entrance examination, English proficiency, and General translation A-B and B-A, in line with Bossé-Andrieu

(1981). The overall aim of our study is to design an instrument that might demonstrate the existence of a hypothetical relationship between reading skills and translation ability, and as such serve as a means of predicting future performance in translation. **Chapters 8 and 9**, which make up **Part II** of this study deal with the broader concerns of research design and methodology, describing the context within which we believe our work is taking place, and delimiting its potential value.

Part III presents the results of our work in seven chapters, each of which offers an analysis of the data obtained from the application of one of the Spanish and English language reading skills sub-tests. In **Chapters 10 to 17** we offer a detailed analysis of the performance of the individual items and of the tests overall, presenting the Facility values and Discrimination indices, and the frequency distributions, and analyzing the results of sub-test administration. We account for the developmental changes made over the period of our study, and attempt to explain the difficulties and shortcomings of our test instruments. In **Chapter 18**, we look in detail at the target parameters we had established for our tests, and describe the levels of correlation achieved.

Finally, in **Chapter 19**, we offer our overall conclusions on the research project and its results.

1.4 Overall research q uestion

We will now present the general and specific research questions on which our study is based, and each of the hypotheses, with the corresponding null hypotheses, that we aim to investigate.

In general terms, the purpose of our study it to establish the adequacy of objectively marked tests in Language A and Language B reading skills, as a means of measuring aptitude for translation.

1.4.1 *HYPOTHESES*

The hypotheses underlying this study, and therefore those that we seek to prove or disprove are that:

H Reading skills provide a measure of aptitude for translation

H_0 No direct relationship exists between reading skills and aptitude for translation

H Translation studies involves a degree of specialization such that a specific entry test of aptitude is justified

H_0 Translation studies does not involve a degree of specialization so great as to warrant the use of a specific entry test of aptitude

H No direct relationship exists between LA reading skills and General translation from LA into LB

H_0 No direct relationship exists between LB reading skills and General translation from LB into LA

H Our LB reading skills sub-tests measure skills and areas of knowledge that form part of general English language proficiency

H_0 Our LB reading skills sub-tests measure skills and areas of knowledge other than those that form part of general English language proficiency

H The introduction of an entry test that includes an LB sub-test will raise the standard of general LB proficiency

H_0 The introduction of an entry test that includes an LB sub-test will have no effect on the standard of general LB proficiency

H The sample of reading skills included in each sub-test is representative of those in our initial test specification

H_0 The sample of reading skills included in each sub-test is not representative of those in our initial test specification

H A specific test of aptitude for translation is not required as the relationships between the OPT scores are adequate measures

H_0 A specific test of aptitude for translation is required as the relationships between the OPT scores are insufficient measures

1.4.2 *THE NULL HYPOTHESIS*

The nature of our study can be expressed through the formulation of the null hypothesis, as an outline of the consequences of committing either a Type I or a Type II error in its resolution, and through the alternative hypothesis (Black 1993; Rowntree 1981). Our null hypothesis, H_{1-0} is that no statistically significant correlations exist between reading skills in LA and/or LB and General translation from English into Spanish, or from Spanish into English.

Table 1—1 The null hypothesis

	Null hypothesis TRUE	*Null hypothesis FALSE*
REJECT Null hypothesis	Type I error	Ok
ACCEPT Null hypothesis	Ok	Type II error

The frame presented below is a typical representation of the Type I and Type II errors, and of the consequences of their (mis-) interpretation. If we apply this, the errors that would occur would be of Type I. This would mean rejecting the null hypothesis, and interpreting correlations as significant, when in fact this was true, and they lacked significance. A Type II error would involve us in accepting the null hypothesis as correct, deeming the correlations not to be significant, when this was false.

The alternative hypothesis was that the correlation would be statistically significant. Then we could conjecture as to the efficiency of the test

1.5 Summary

Several thousand candidates take entry tests in aptitude for translation at universities across Spain each year, yet we have found virtually no empirical research on the examinations — for which some universities oblige candidates to pay — although the first research in progress is beginning to appear (Fox 1997). This alone seems sufficient justification for our study.

In 1994, the University of Granada decided to charge candidates an examination fee in order to take the test, making this one more of the many examinations belonging to the language testing industry. Such a decision demands of those who design and implement the test the maximum rigor in its design and administration, and in the interpretation of its results. Sadly, this source of income has not yet been used to finance research into aptitude for translation.

One significant aspect of our study may lie in the four specific areas under consideration, each of which was crucial to its conception and development. These are the theoretical, operational, and ethical considerations arising from testing for "aptitude for translation", and the logistical considerations pertaining to the situation in the University of Granada.

Testing for aptitude has a built-in flaw, in that any work on subsequent measures involves at best all of those who have "passed" the initial test. Those who fail are no longer available for the follow-up study. At worst, this work can only be carried out on a sample of these. In addition to this constraint, the longitudinal nature of our research has introduced a number of variables beyond reasonable control. While Bossé-Andrieu (1981) reported a similar study in which she did not discuss either the time factor, or the other intervening variables, we feel that in our study we cannot ignore these. Such variables have a clear impact on those elements of analysis purporting to reflect on the validity of the initial hypothesis. However, conclusions as to the essential efficiency of the test as a measure of initial selection cannot be said to have been unduly influenced, and would appear to stand up to reasonable analysis.

Our study affirms that aptitude for translation can be measured. We approach the matter of the test instrument with the hypothesis that reading skills are a crucial part of the translation process, specifically of the pre-translation process, and that the translator must be an "expert reader". Given the constraints established in the legal parameters of the test, we believe these skills can serve as a basis that does not presuppose any specific preparation in Translation studies on the part of the candidate, and we believe we are able to demonstrate this.

The longitudinal part of the study sets out to test this hypothesis by correlating the independent variables of LA and LB reading skills against the dependant variable of General Translation from English into Spanish (A-B) and from Spanish into English (B-A). We do not aim to produce more than a description of the strength and nature of any such relationships. The cross-sectional part of the study is designed to produce a list of the specific reading skills and the test task type(s) which best serve to discriminate between those candidates who demonstrated a greater or lesser degree of aptitude.

In our results, we have found weak, statistically significant and non-significant, positive and negative correlations between the University entrance examination, and Secondary school averages, and the examinations in General translation A-B, and General translation B-A. These suggest that Translation studies does involve a degree of specialization sufficient to justify a specific entry test of aptitude.

We have also gathered comparative data that indicate that the introduction of an entry test is fully justified, and that it had no negative effect on the standards of general English proficiency among students entering the program. What is more, the data suggest that there has been a gradual improvement in standards, although we can be certain this is not due to the entry tests alone. Teachers had commented on just such a trend, but before this study, no data had been gathered to support it.

It would be too much to expect that an in-house study such as ours could possible achieve results that improved on the many, many well-financed aptitude research projects documented in the history of testing (Spolsky 1995). However, it is of much comfort to discover that the range of results in terms of correlation coefficients, is encompassed by that which generally appears in the literature ($r = 0.20$ to 0.60). At the very least, we have done "as well" as our predecessors.

2 Annotated definitions

2 ANNOTATED DEFINITIONS

2.1 Analytic scoring
2.2 Aptitude
2.3 Aptitude test
2.4 Battery
2.5 CEGEPs
2.6 Discrete item
2.7 Discrimination index (DI)
2.8 Equivalence
2.9 Examination grades.
2.10 Face validity
2.11 Facility values
2.12 Frequency distribution
2.13 Global assessment
2.14 Guttman split-half coefficient
2.15 Index Score
2.16 Integrative item/task
2.17 Inter-rater reliability
2.18 Kendall's *tau*$_B$
2.19 Key response
2.20 Kuder-Richardson reliability coefficients.
2.21 Language A
2.22 Language categories
2.23 Moderation
2.24 Parallel test reliability
2.25 Pearson product-moment coefficient
2.26 Rater
2.27 Reliability
2.28 Sampling error mean
2.29 Skewness
2.30 Skill areas
2.31 Spearman-Brown split-half coefficient
2.32 Standard error of measurement
2.33 Sub-test
2.34 Weight
2.35 Z-values

THIS CHAPTER is intended as an aid to the reader. We are well aware that for some it is unnecessary, and for others less than adequate. We can only hope that for all it may serve some purpose, if only in providing a selection of bibliographic references to further research.

2.1 Analytic scoring

> *A method of scoring which can be used in tests of productive language use, such as speaking and writing. The assessor makes an assessment with the aid of a list of specific points. (ALTE 1998:135)*

Analytic scoring is the method used in testing methods such as the Summary writing task that is traditionally used to test aptitude for translation.

2.2 Aptitude

> *...a* concept-*a construct, if you will-referring to some constellation of* conditions, *presumably residing in the individual, that predispose him to either success or failure (or some point along the continuum between these poles) in some future activity, in particular some activity requiring new learning. (Carroll, J.B. 1974:286)*

2.3 Aptitude test

> *A test designed to predict or measure a candidate's potential for success within a particular area of learning, e.g. in learning a foreign language, or on a specific course of study. (ALTE 1998:135)*

2.4 Battery

> *A set of related tests or sub-tests which make independent contributions (e.g. by testing different skills) but may be combined to produce a total score. (ALTE 1998:136)*

In the case of the MLAT (Carroll and Sapon 1959) and the OPT (Allan 1990) a series of sub-tests have been designed to cover what the authors considered to be the range of skills involved in FL aptitude. In out research, we have a test specification that includes three sub-tests: reading skills in Language A, reading skills in Language B, and listening skills in Language B.

2.5 CEGEPs

Collège d'enseignement général et professionnel. In the Canadian education system, these centers are similar to the 6th Form Colleges found in England and Wales, but they cover a wider range of subjects including professional and vocational studies.

2.6 Discrete item

> *A self-contained item. It is not linked to a text, other items or any supplementary material. (ALTE 1998:142)*

The task types we use in these sub-tests are not discrete items, as they are always based on the use of language in a text.

2.7 Discrimination ind ex (DI)

The value of a specific item as a means of discriminating between those candidates who achieve higher scores, and those who achieve lower scores (Heaton 1988:179-180) is recorded on a scale from -1 to +1, and known as the Discrimination index. The population is rank-ordered by scores from highest to lowest. Those candidates who fall within the first 27.5% are labeled 'U', and those who fall within the last 27.5% are labeled 'L'.

The DI score of an item is calculated with the following formula:

$$DI = \frac{(u-l)}{27.5\%*N}$$

u represents the number of correct responses among the U candidates; l represents the number of correct responses among the L candidates; N represents the total number of candidates.

2.8 Equivalence

> *To meet the strict requirements of equivalence ... different forms of a test must have the same mean difficulty, variance and covariance when administered to the same persons. Equivalence is very difficult to achieve in practice. (ALTE 1998:144)*

2.9 Examination grad es.

Final grades are awarded against an ordered category scale (Rowntree 1991:29-34). This means that the individual students are classified on a "better or worse" basis: distinctions are relative, and they are not made individually, but for groups of students. The numbers derived from arithmetical scores do not indicate qualitative differences: a student who scores 7.0 is not twice as "good" as one who scores 3.5. The five categories that we refer to in this study, and the arithmetical ranges to which we equate these, are Matricula *de honor* (10-9.5); *Sobresaliente* (9.4-8.0); *Notable* (7.9-6.5); *Aprobado* (6.4-5.0); *Suspenso* (4.9-0). We do this by following the "*Tabla de equivalencias de notas: EUTI-AIX-ECHE*" (1987). This is an in-house working document published for the purposes of the Applied Languages Europe (ALE) program. It was written by the International course Committee ALE, based on a statistical study of marks at Aix and Ealing College of Higher Education (now Thames Valley University, London) over several years. After a "rather brief" study of marks in Granada the table was adapted to include these in 1989. At the time of writing a further modification is being considered (Personal communication Dorothy Kelly, Coordinator, Applied Languages Europe, Granada 1993.).

2.10 Face validity

This is a subjective judgement as to whether or not a test, superficially, appears to the candidates or others to be an acceptable measure of whatever it purports to be testing. (Bachman 1990:285-289.)

2.11 Facility values

The proportion of correct responses to an item, expressed on a scale of 0 to 1. (ALTE 1998:145)

The Facility value (FV) is one of the two basic instruments of item analysis that enable us to assess the performance of individual items, and thus compare the range of items in a test.

2.12 Frequency distribu tion

In test data, the number of occurrences of each score achieved by candidates. (ALTE 1998:143).

In our discussion of results, we use a histogram to demonstrate the nature of specific distributions in line with standard practice.

2.13 Global assessment

> *A method of scoring which can be used in tests of writing and speaking. The assessor gives a single mark according to the general impression make by the language produced, rather than by breaking it down into a number of marks for various aspects of language use. (ALTE 1998:146)*

Writers such as Heaton (1988) and Renfer (1992), make use of the opposition between an ANALYTIC and a GLOBAL approach to test marking. They use the former to describe a detailed evaluation of specific, usually predetermined aspects; the latter describes a more impressionistic method of marking a test. Heaton (1988) also describes a third method: the mechanical accuracy, or error-count method, which he considers the least valid approach as it concentrates only on the negative aspects of student performance. Global marking in particular can give rise to discrepancies in both intra- and inter-rater reliability.

2.14 Guttman split-hal f coefficient

Formula (Bachman 1990:175):

$rxx^{1} = 2(1\text{-}((s^{2}_{h1} + s^{2}_{h2})/\ s^{2}_{x}))$

We use this formula as one of the means of calculating the internal consistency measure of reliability.

2.15 Index Score

In the context of the MLAT (Carroll John B. and Sapon 1959), this is a scaled score used at the American Foreign Services Institute (FSI) and based on the total score drawn from the MLAT sub-tests. The original mean was 50 with a standard deviation of 10. These norms are today outdated and currently a simple conversion process is used. Learners at the FSI were among those used by Carroll and Sapon when originally norming the performance of the MLAT (Ehrman 1995a).

2.16 Integrative item/ task

> *Used to refer to items of tasks which require more than one skill or subskill for their completion. Examples are the items in a cloze test, an oral interview, reading a letter and writing a response to it. (ALTE 1998:148)*

The Summary exercise traditionally used to test aptitude for translation is an example of an integrative task.

2.17 Inter-rater reliabi lity

> *The degree of agreement between two or more assessors on the same sample of performance. (ALTE 1998:148).*

This aspect of reliability is most difficult to achieve in the marking of subjective tasks such as the summary writing activity commonly used in entry tests to translation courses.

2.18 Kendall's tau_b

Kendall's tau_b (Kendall 1975) measures the number of single step changes required to convert one rank order into another. Normally, it compares the natural rank order, i.e. 1, 2, 3, ... n with an order derived from arithmetic scores. The standard procedure when dealing with tied cases, that is instances when more than one candidate has achieved the same score or grade, is to divide the sum of the rank order positions occupied by the tied individuals by the number of individuals. Each of these is then awarded the resulting rank order position. The next position in the rank order then follows on in the natural order. This example will clarify the procedure:

Worked example of Kendall's tau_b

Case	*Raw score*	*Rank order*	*(Natural order)*
A	6,85	1	1
B	6,70	2,5	2
C	6,70	2,5	3
D	6,47	4	4
E	6,32	5	5
F	5,53	7	6
G	5,53	7	7
H	5,53	7	8
I	5,40	9	9
J	5,15	10	10

The rank order positions of 2.5 and 7 are judged on the basis of these calculations:

$$\frac{\text{Natural order ranks}}{\text{N}^\circ\text{ of tied cases}} \qquad \frac{2 + 3}{2} = 2.5$$

$$\frac{\text{Natural order ranks}}{\text{N}^\circ\text{ of tied cases}} \qquad \frac{6 + 7 + 8}{3} = 7$$

The procedure demonstrated above was used in all cases when handling interval data because the number of tied cases, and the numbers if individuals in each instance was small. Ordinal data was predicted to produce four tied groups in every instance corresponding to the four standard ordinal ranks: Fail, Pass, Credit, and Merit. Our aim was that the sample number should be as large as possible (min ≥ 30) so we knew this would cause a loss of information. In order to generate manageable data we assigned each of the four ranks a numerical score:

Suspenso	4
Aprobado	5
Notable	6
Sobresaliente	7

On which basis the rank order coefficient could be accurately calculated.

As Siegel and Castellan (1988:241) indicate:

> *... the effect of ties in the ranking is to inflate the value of the (uncorrected) correlation [...], for this reason, the correction should be used where there is a large proportion of ties [...] or the number of ties in a grouping is large.*

Both of these conditions pertained to our study.

We were able to analyze our data to a greater level of sensitivity by using the tau_b procedure (Kendall 1975:34-48; Siegel & Castellan 1988:239-241; 249-251). Once we had established the tau_b coefficient, we could calculate the value of z, and then ascertain the level of significance of our results. The table of probabilities associated with the upper tail of a normal distribution (Siegel & Castellan 1988:319) for groups such as ours, were $n \geq 30$, provided this data. The formula required to calculate z was

$$z = \frac{3T\sqrt{N(N-1)}}{2\sqrt{(2N+5)}}$$

2.19 Key response

The correct option in a multiple-choice item. (ALTE 1998:150).

2.20 Kuder-Richardson reliability coefficients.

KR-20

$$r_{tt} = \frac{n}{(n-1)} \left(\frac{s_t^2 \; s_i^2}{s_t^2} \right.$$

In this formula r_{tt} represents KR-20; n = the number of item; s_t^2· the variance of the test scores; s_i^2, represents the sum of the variances of all of the items.

KR-21

$$r_{tt} = \frac{n}{(n-1)} \; 1 \qquad \frac{x \, x^2/n}{s_t^2}$$

In this formula r_{tt} represents KR-21; n = the number of item; s_t^2· the variance of the test scores; x represents the mean of scores on the test.

2.21 Language A

In the literature on language teaching and learning three terms are regularly applied in order to distinguish the different languages that an individual uses: native language, second language (SL) and foreign language (FL) (Stern 1983:15-18). However, in line with general practice in Translation studies, we adopt the International Association of Conference Interpreters' definitions (AIIC 1982:10). These do not distinguish between native language and language of habitual use. The terms we will use are Language A (LA) to denote Spanish, defined as the language in which translators and interpreters possess native-like proficiency. Secondly we will talk about LANGUAGE(S) B (LB), for the language(s) which candidates are supposed to master both actively and passively almost as well as a native, but which are active working languages, meaning that they are supposed to work into them. Finally, we will refer to LANGUAGE(S) C (LC), which are passive languages. They are theoretically understood at native level, but students are not taught to work into them. At the University of Granada, the degree studies program assumes Spanish to be the LA of all candidates. Candidates have one LB, chosen from English, French and German. In addition, they study LCs — such as Arabic, Russian, Chinese, Portuguese or Italian — that are learned from *ab initio* level. These include all other languages taught within the university, although a distinction is made between LC Is, which students learn to translate into LA, and LC IIs, most notably Dutch, which they do not learn to translate into or from.

2.22 Language categories

The US Foreign Service Institute divides languages into four categories according to their difficulty, which is measured as a factor of the length of time NS English learners take to achieve specified levels of attainment (Ehrman 1995a:6). The categories are as follows:

Category 1 languages: Western European

Category 2 languages: non-western European languages which are "relatively quick" for NS English learners e.g. Swahili, Indonesian, some North European languages

Category 3 languages: Eastern European and non-western languages (Except those in Category 4) e.g. Russian, Thai.

Category 4 languages: "Super-hard" languages e.g. Arabic, Chinese, Japanese, Korean

2.23 Moderation

When marking candidates' scripts, moderation takes place if one examiner adjusts the marks of other examiners to ensure that all have been using the same criteria, or applying the markscheme in the same way.

2.24 Parallel test reliability

In practice, this is a concept that is very difficult to achieve. If two tests have been written according to the same specification and are administered to candidates belonging to the same population, under the same or similar conditions, then the results obtained should be the same. In order to establish the equivalence between two forms of the same test we need to calculate the mean difficulty, variance, and covariance when administered to the same candidates (ALTE 1998:144). In our case, we are not able to establish parallel test reliability for the sub-tests es01.2 and es02 because they were administered to two different groups.

2.25 Pearson product-moment coefficient

This formula calculates the relationship between two sets of scores achieved by the same individuals. The formula is based on equating the raw scores by converting them to Z-VALUES, that is units of STANDARD DEVIATION. The formula used is

$$r_{xy} = \frac{z_x z_y}{N-1}$$

N = the number of individuals in the sample. (Henning 1987:59)

2.26 Rater

This is the name given to someone who assigns a score to a candidate's performance in a test, using subjective judgement to do so. Raters are normally qualified in the relevant field, and are usually required to undergo a process of training and standardization.

2.27 Reliability

The consistency or stability of the measures from a test. The more reliable a test is, the less random error it contains. A test which contains systematic error, e.g. bias against a certain group, may be reliable, but not valid. (ALTE 1998:160)

2.28 Sampling error mean

For some samples the (SE_m) was calculated in order to establish the validity of results. The formula used was:

$$\frac{SD_{pop}}{N} = \underline{Sd_{test}} = SE_m$$

N = N° of respondents

test = the specific test

2.29 Skewness

This term describes a statistical distribution that is not symmetrical. The *skew* is the long tail of observations that may go to the right of the curve, or to the left. A right-side skew is sometimes described as being positive and a left-side skew as negative (Rowntree 1981:58-59). When we look at the frequency distribution of candidates' scores on an examination paper, a left side or negative skew indicates the paper was relatively easy: more candidates scored higher marks; a right side or positive skew indicates the paper was difficult.

2.30 Skill areas

In discussing the MLAT, the US Foreign Service Institute distinguishes between an S-rating, for speaking and interactive skills, and an R-rating, for reading.

2.31 Spearman-Brown split-half coefficient

Formula (Bachman 1990:175):

$rxx^1 = 2rhh^1/(1+rhh^1)$

We use this formula as one of the means of calculating the internal consistency measure of reliability.

2.32 Standard error of measurement

> *...the standard error of measurement (S_e) is an indication of the imprecision of a measurement.* (ALTE 1998:164)

To calculate S_e we use a formula based the reliability (r) and the standard deviation of the test scores (Sx).

$S_e = Sx\sqrt{(1-r)}$

We interpret candidate performance by adding or subtracting the S_e to the observed score. The degree of confidence with which we can judge scores increases by steps as it does when interpret a NORMAL DISTRIBUTION: 68% represents S_e, 95% represents $2S_e$.

2.33 Sub-test

> *Part of an examination often presented as a separate test...often skills-based. (ALTE 1998:139)*

Our entry test is comprised of three sub-tests, two of which measure reading skills, in LA and LB, and the third of which measures LB listening skills.

2.34 Weight

> *The assignment of a different number of maximum points to a test item, task or component in order to change its relative contribution in relation to other parts of the same test. (ALTE 1998:169).*

In our research, the weighting of components occurs in sub-tests in03 and es04. The Error identification and Error correction items of these tests, which total 20 raw points each, are weighted equally with the 10 raw points of the MCQ items, giving a final total out of 20, and not 30.

2.35 Z-values

These are typically used to compare scores from two different variables. By subtracting the mean from the individual's score, and then dividing by the standard deviation (SD), we can calculate the z-value for any one individual. Z-values will always have a mean of 0 and a SD of 1 (Rowntree 1981:80-81; ALTE 1998:169).

Part I
Background

3 CONTEXT

3 CONTEXT

3.1 AN OVERVIEW OF THE CONTEXT
3.2 A NEW OPPORTUNITY
3.3 PUBLIC DEBATE OVER PROGRAMS
3.4 THE DIPLOMA EXPERIENCE: AT THE GRASS ROOTS LEVEL IN GRANADA
3.4.1 *LB English failure rates*
3.4.1.1 The context: Traducción I (Inglés) as taught in the Diploma
3.4.1.2 Student numbers
3.4.2 *Pre-diploma course LB English proficiency*
3.4.2.1 A first attempt at damage limitation: 1991-92
3.4.2.2 Discussion
3.4.2.3 A second attempt: 1992-93
3.4.2.4 Conclusion
3.4.3 *The synthetic approach to aptitude testing in Spain*
3.4.3.1 The Diploma course entry test for mature students, Granada April 1992
3.4.3.2 Discussion
3.5 PREDICTED CANDIDATE NUMBERS
3.6 ADMISSION TO TRANSLATION STUDIES AROUND THE WORLD
3.6.1 *Bossé-Andrieu's study of francophone countries*
3.6.1.1 Discussion
3.6.1.2 Conclusions
3.6.2 *A parallel study: Campagna and Dionne (1981)*
3.6.3 *A postscript from western Canada*
3.6.4 *Attitudes in the US*
3.6.5 *A worldwide perspective*
3.6.5.1 Austria
3.6.5.2 Uruguay
3.6.5.3 Israel
3.6.5.4 Switzerland: Zurich School for Translation and Interpretation
3.6.6 *Japan: The question of vocational aptitude among professional translators*
3.6.7 *Survey of findings*
3.7 CONCLUSIONS

THE OBJECTIVES of our review of the immediate context in which the need to develop an initial entry test for Translation within the degree program in Translation and Interpreting were

- To identify a model of test and test specification that could be applied in the University of Granada
- To describe the underlying theoretical principles behind such a test, or those which might be applied in the preparation of a test
- To identify a research design model that might be used in the process of developing and refining an entry test.

3.1 An overview of th e context

In the first part of this chapter, we describe the historical context in which the need for an aptitude test arose, beginning with an outline of some aspects of the background against which the degree program was developed. We then go on to present a detailed account of relevant aspects of teaching on the Diploma in Translation and Interpreting at the University School of Translators and Interpreters (*E.U.T.I*) of the University of Granada in the years immediately preceding the first aptitude test. Here we include a description of the de facto "entry" test provided by the first year Diploma subject *Traducción I Inglés*. We draw attention to its drastic consequences in terms of the student failure rate, and we analyze two attempts to pre-test student proficiency in English, their first foreign language, or LANGUAGE B (LB), in order to remedy this situation. To do so, we compare the results of a pre-course proficiency test prepared in-house, with those of an objective assessment of student ability, measured by an externally validated test of proven reliability (Allan 1992). The results of this second test will later serve as one of the benchmarks by which we attempt to validate our aptitude test.

In the Diploma course the entry procedure for mature students was the only area of university studies that permitted a specific test of aptitude. And, in a description of the model of aptitude test used for these candidates, we will highlight some of the pitfalls of the uncoordinated use of a synthetic or integrative test, which relies for its success on a consensus over criteria and methods of marking and grading. Finally, we present data for Diploma course applications, on the basis of which "messy" logistical considerations — in terms of applicant numbers, and of the administrative timescale involved — demonstrate themselves to be fundamental to the test design process

From the immediate context of the Spanish Diploma course and the introduction of degree level studies, we broaden our sights.

We now survey research into entry procedures and pre-course selection, and into the measurement of aptitude around the world in order to fulfil our three objectives. Principally, we describe work carried out from Canada and dealing with the situation in francophone countries (Bossé-Andrieu 1981; Campagna & Dionne 1981; Rainey 1988). However, we also include reports on procedures in other parts of the world (Weber 1984; Snell-Hornby 1992; Sainz 1992; Amit-Kochavi 1992; Renfer 1992), and on the vocational aptitude of professional translators (Szuki 1988).

In our conclusions, we indicate our failure fully to achieve any of these aims. We note the divided opinions about the relative importance of three aspects of test procedures. The first of these is the matter of the balance between the LA and the LB as the focus of the test procedures. The second is the nature of the test mechanisms to be used, whether these should be synthetic or analytic. And the third is the question of the language skills to focus on, and whether the so-called active skills of written expression, or the passive skills of reading comprehension, should take precedence. Perhaps the only success is in terms of the research design used by Bossé-Andrieu (1981) and its possible application to our study.

3.2 A new opportunity

The creation of a first degree program in Translation and Interpreting was a major step forward in the academic and technical preparation of future members of our profession. The increase from a three-year diploma to a four-year degree course, and the added academic demands implied would provide opportunities to improve the depth and breadth of preparation. It would also be an opportunity to establish minimal entry requirements, over and above those of the University entrance examination – *selectividad* – which would ensure a high level of academic ability among students. The debate in the immediate context of the establishment of the degree program developed in two areas: the politico-academic discussion over contents, programs, and areas of academic influence; and the grass roots arguments over standards of LB proficiency.

3.3 Public debate ove r programs

In 1988 the proposed degree course program was published in a discussion document that included four alternative versions and more than twenty other comments, observations or suggestions from university departments, professional associations, and private individuals (*Consejo de Universidades* 1988). In the formal proposal under debate reference appeared to an entry test of aptitude -*una prueba de aptitud inicial* - which should be considered - *debería contemplarse* - for inclusion in the regulations of the degree (21). The importance of this test was heavily underlined by other contributors too (58; 106), although no one, at any point raised the question of the nature of aptitude for translation, nor the matter of how it could be measured.

The attitude of translation professionals reveals the status of LB teaching and learning within Translation studies. Spain's professional association of translators and interpreters (*Asociación Profesional Española de Traductores e Intérpretes,* APETI) criticized the teaching of the diploma course. The basis of their view was that it had been incorrectly based on the model of modern language teaching, rather than on the training of professional translators and interpreters (135-41). The question of the primacy of LB proficiency, at least in the minds of the University teachers and administrators involved, was challenged: a primacy, which had not been demonstrated to have any empirical base.

The wider debate on curriculum design and development was the focus of a seminar held in Granada later in the same year (*I.C.E.* 1988). Specialists from the University of Copenhagen, with much experience in course programming, directed the work of the seminar. Surprisingly, the question of an entry test did not even appear in the published conclusions.

3.4 The Diploma expe rience: at the grass roots level in Granada

Although no detailed, empirical study had been published on the topic, the experience much commented at anecdotal level within the University of Granada was that LB English proficiency on entry to the diploma program had been unsatisfactory. We quantified this experience to provide benchmark standards against which to measure our test. Data were taken from statistical records facilitated by teachers and the Faculty office. We focussed on two aspects: firstly, the high failure rate experienced in the Diploma course subject *Traducción I (Inglés)*, half of which — *traducción inversa* — was partly equivalent to the degree course subject LB English. Secondly, on LB English proficiency as measured in pre-course tests. The data referred to academic years 1991-92 and 1992-93, the two years immediately before the introduction of the degree program.

3.4.1 *LB ENGLISH FAILURE RA TES*

Concern among university teachers had primarily been over LB English proficiency. Given that any entry test had been judged unconstitutional, this led to the use of first year studies as the a posteriori process of selection, bringing about an exceptionally high failure rate. This paralleled the situation in some Belgian centers and in Austria (Bossé-Andrieu 1981; Snell-Hornby 1992; Argüeso 1994).

3.4.1.1 *The context:* Traducción I (Inglés) *as taught in the Diploma*

Before presenting the details, the context from which our data were drawn needs to be clarified both in terms of the subject itself, and in terms of the student numbers concerned. While *Traducción I (Inglés)* was taught and examined in two parts *traducción directa* and advanced English reading and writing skills — a number of different evaluation exercises were involved (I.C.E. 1991:94-101). These included two course final examinations in written production: the translation of a journalistic text from English into Spanish, and an essay in English. In 1992-93, in an attempt to improve the validity of the evaluation exercises, a summary writing exercise was also used: candidates read a text in English and wrote a 50-word summary, also in English (Newmark 1988). The other assessment activities carried out over the year included an exercise in proof-correction, a short translation task to be handed in at the end of the course, note-taking and formal essay planning.

The teaching program for the English language skills half of the subject laid emphasis on reading skills and on a process-oriented approach to skill development. The emphasis in evaluation was on productive tasks, and there was an analytic approach taken to the teaching and testing of skills perceived to be basic to translation. Materials drawn from two areas formed the base of the course. Firstly, there was a range of theoretical texts focussing on reading skills and on the role of the reader (Booth 1961; Iser 1974; Traugott and Pratt 1980; Grellet 1981). Resource materials taken from practical, English as a Foreign Language (EFL) course books complemented these (Moore, Fernanda de Knight, Munévar, and Bonnet de Salgado 1979; Barr, Clegg, and Wallace 1981; Rudska, Channell, Putrseys, and Ostyn 1981; Hoey 1983; Hess 1991; Holme 1991).

Although the official program made no mention of the fact, the course final examination was WEIGHTED. Ostensibly, candidates had to pass the two halves of the subject, and equal importance was given to each of these. In fact, the balance lay heavily on the LB skills, as this examination was eliminatory. All candidates had to pass the English language essay examination *before* their B-A translation papers were even marked. This ensured that the Pass/Fail boundary for LB English provided a selection procedure based on the individual student's language abilities in reading and writing skills. The scale of the failures resulting from this policy was all too evident, and we present this graphically in Figures 3—1 and 3—2.

For the essay writing component, the teachers involved subjectively marked the end-of-year examinations, they did not

Figure 3—1 *Traducción I (Inglés)* 1991-92 June Session The high failure rate among 1st year Diploma students, and their "fear" of failing are clearly shown in this chart

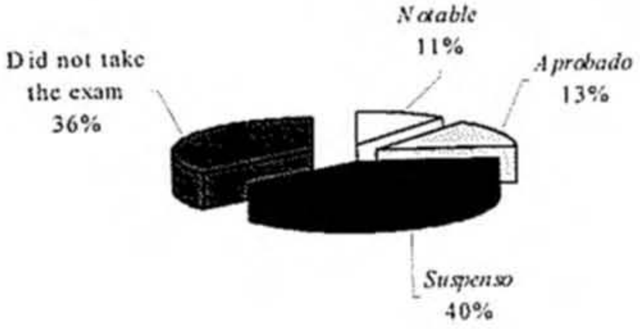
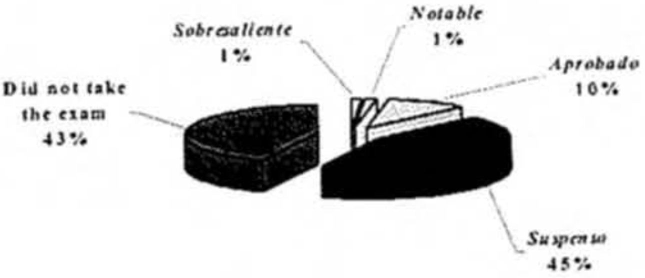


Figure 3—2 *Traducción I (Inglés)* 1992-93 June session



carry out any formal intra- or inter-rater reliability checks. This made it impossible to demonstrate that they had been consistent in the criteria they applied throughout the batch of examinations they had marked, or that they had all used the same criteria. The members of the Examining board controlled the question of evaluation criteria on a purely ad hoc basis, as and when problems arose during examination sessions. In general the kind of problem that provoked discussion and subsequent decision-making was to do with students who had failed the examination on five separate occasions, and were therefore only able to sit it once more. In response to the vagaries of this situation, at the beginning of the 1992-93 academic year, a set of criterion descriptors for written production was published. At the same time, the achievement level for the course was set at equivalent to Cambridge Advanced English in reading and writing skills, and this was also made public.

3.4.1.2 *Student numbers*

From the data we encountered we can only conclude that the subject in question, *Traducción I (Inglés)* clearly proved an efficient pseudo-entry test. Our analysis of the failure rate makes this quite clear. Access to the Diploma course was limited to 200 students per year, divided into three divisions: 80 students were admitted to study the Spanish-English combination, 80 more for Spanish-French, with 40 admitted for Spanish-German.

In 1991-92, 218 students registered for *Traducción I (Inglés)*; a year later there were 221. All were eligible to take the exam in the corresponding sessions, but in June 1992 only 140 of them (64%) actually decided to sit the exam. A year later this figure fell to 126 (57%). The majority of these students had either failed, or chosen not to take the examination in previous sessions. Many, when asked, would say they had decided not to take the examination previously for fear of failing. In the two sessions in question, the failure rate, (Figures 3—1 and 3—2) was as high. In 1992 this reached 63% (88 of 140 students registered), and in 1993 it rose to 78% (98 of 126 candidates) (*Actas de Traducción I (Inglés)* 28 June 1993; 8 July 1992).

3.4.2 *PRE-DIPLOMA COURSE* **L B** *ENGLISH PROFICIENCY*

The application of a de facto entry test through the first year subject *Traducción I (Inglés)* indicates the level of concern among teachers over the standard of LB English among students on the diploma course. It is not our concern to question the motives behind their preoccupation, but it is clear that some aspects of the student population were less than satisfactory. The very wide range in levels of English proficiency and the class sizes were in themselves factors that would inevitably impede the learning process.

3.4.2.1 *A first attempt at damage limitation: 1991-92*

A first attempt at damage limitation was initiated in the academic year 1991-92 when seventy-six first year Diploma course students were pre-tested to ascertain their level of proficiency. In October 1991 a test loosely based on a *Cambridge Proficiency in English* past paper (date unknown) was conducted, and three members of staff marked the scripts. As in the case of the course final examinations for *Traducción I (Inglés)*, neither intra- nor inter-rater reliability were checked, nor was the test externally validated. The resulting frequency distribution of the scores attained by candidates, shown in Figure 3—3 below, and the descriptive statistics calculated from their results revealed the inadequacy of the test instrument despite the best intentions that lay behind its use.

The frequency distribution of a set of test scores will reflect certain aspects of the test itself, and of the individuals who take the test. The larger the sample, the more the curve can be expected to resemble the NORMAL DISTRIBUTION represented by a bell-shape (Rowntree 1981:57-81). The frequency distribution derived from this test was far from "normal". It did not have two equivalent peaks, but the existence of two separate curves was obvious. The first of these had a negative skew based around the mode of 10 out of 20; the second had a positive skew based around the mode of 14 out of 20. These curves indicated that the test instrument was measuring two separate populations. If the purpose of the test had been to divide students into two broad bands according to ability it would have more than satisfactorily served its purpose. As can be seen (Figure 3—3), a rough break could be made between 12/20 and 13/20, representing a trough between the two curves. Furthermore, a range of 16 marks out of a possible 20 is clearly strange.

3.4.2.2 *Discussion*

We conclude that the raters used different criteria in the marking process. The testing exercise was basically flawed as no double marking of papers or any other form of moderation of criteria or scores took place. These errors within the marking process will later be seen to be repeated when we describe the entry test for mature students held in April of the same academic year. One positive aspect of this exercise was the fact that numerical marks were recorded for all candidates enabling us to carry out the statistical analysis shown here.

3.4.2.3 *A second attempt: 1992-93*

A change of approach followed in 1992-93, when a commercially available test instrument - the *Oxford Placement Test* (OPT) (Allan 1990) - was administered as the pre-course test. This test was chosen because of the detailed information about its reliability and validity published in the manual. It was also seen to be relatively easy to administer and correct. A further point in its favor was the fact that scores are calibrated against a set of proficiency levels expressed in terms of widely recognized examinations, as shown in Table 3—1.

Table 3—1 English language proficiency levels as described in the OPT, by reference to other criteria. Adapted from Allan (1990).

Level	*Descriptor*
Advanced to near native	Good Cambridge CPE level and above. Potential candidate for Cambridge Diploma.
Upper intermediate to advanced	Potential candidate for Cambridge CPE, ARELS Diploma, RSA Stage III, RSA Communicative Test at advanced level.
Intermediate to upper intermediate	Potential candidate for Cambridge FCE, ARELS Certificate, RSA Stage II, RSA Communicative Test at intermediate level, Oxford Delegacy Higher.
Lower intermediate to intermediate.	Potential candidate for Cambridge PET, ARELS Prelim, RSA Stage I, RSA Communicative Test at basic level, Oxford Delegacy Preliminary.
Post-elementary to lower intermediate	
Elementary to post-elementary	

The full OPT Listening sub-test was used in combination with the Grammar Part One sub-test. The Grammar Part Two sub-test was not used as the OPT instructions (Allan 1990) suggest this is best reserved for discriminating at more advanced levels. The scores for the two sub-tests were given equal weighting, and the results, scaled to a score out of 20, presented a fairly normal frequency distribution (Figure 3—4) and acceptable descriptive statistics (Table 3—2). Ninety-four percent of students were at or below the level of Cambridge First Certificate, and only 9.6% were already at the target level established for the course.

Figure 3—3 *Traducción I (Inglés)* Pre-course proficiency test 1991-92. What a frequency distribution should *not* look like: the "twin peaks" clearly show something was wrong with this test.

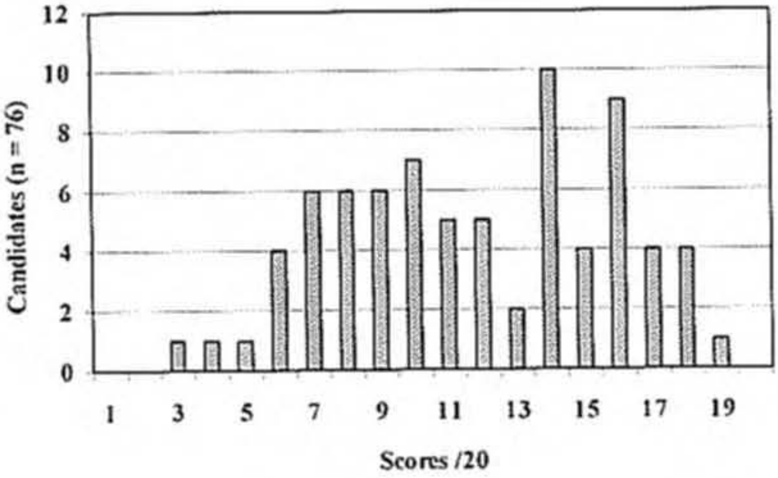


Figure 3—4 *Traducción I (Inglés)* Pre-course proficiency test 1992-93. The curve produced by these results is much more "bell-like", than that of the previous year's test.

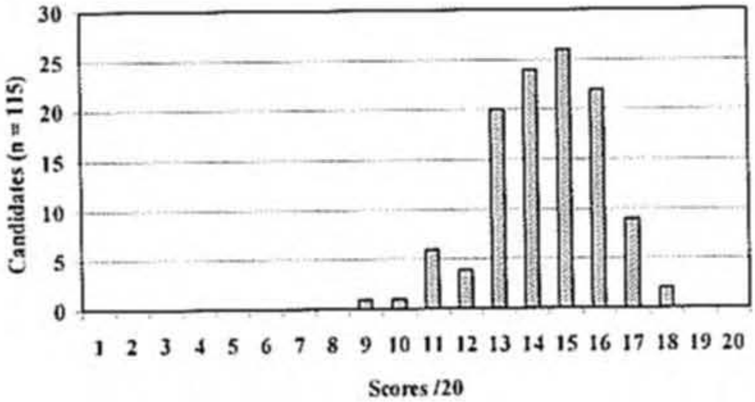


Table 3—2 Pre-course proficiency test 1992-93. An acceptable set of descriptive statistics.

Mean	14.04
Median	14.10
Mode	15.5
Standard error	0.16
Standard deviation	1.72
Skewness	(0.48)
Variance	2.97
Level of confidence: 95%	0.31

A comparison of the two histograms clearly highlights the contrast between the different pre-course tests. This difference could be explained in two ways: either, the commercially produced testing instrument used in the second year produced much more accurate figures, or the nature of the population changed remarkably from one year to the next. We are inclined to favor the former explanation.

3.4.2.4 *Conclusion*

From our analysis of the situation regarding LB English proficiency, and the attempts made both to establish minimum standards and to measure proficiency levels, we concluded that the previously undocumented supposition that an entry test was essential did have an empirical foundation. The process of analysis we have described enabled us to establish the OPT as one instrument that could later serve as a benchmark against which to validate our proposed test, at least in as much as this could be said to measure LB English proficiency. Consequently, we can say that the Diploma experience would serve as a means of gauging whether or not our test, at the very least, produced results that did not make the situation worse.

3.4.3 *THE SYNTHETIC APPROACH TO APTITUDE TESTING IN SPAIN*

However, one area of university entry procedures did permit the use of an aptitude test, and that was within the regulations for the admission of mature students— those over 25 years of age. This gave us the opportunity to study a specific model of test and consider its value for further use in the degree course entry process. The exercise that we now describe contains a number of the essential flaws we have previously seen in our analysis of the diploma pre-course proficiency test. In this case, these are compounded by a lack of any numerical marking that might be susceptible to statistical analysis.

3.4.3.1 *The Diploma course entry test for mature students, Granada April 1992*

Before the degree program, the only specific entry test that could be used was one designed to select suitable candidates from among the mature students who applied to enter the University School of Translation and Interpreting (EUTI). In April 1992, no written description of this test instrument existed, but the test "specification", such as it was, was passed on by word of mouth from one teacher to another. The test, essentially of the synthetic type, demonstrated an emphasis on the quality of written performance in LB, which had been central to the concerns of teachers. This model is still used in one form or another by many centers teaching translation in Spain, and in Chapter 9, we describe an improved version currently in use at the Autonomous University of Barcelona (UAB). For the moment, we will describe the test as administered in Granada.

Three members of the teaching staff participated in an Examination board with full, autonomous authority for all aspects of the test. They were chosen to represent each of the three LBs, English, French, and German. A fourth member of staff, a native speaker of Spanish was invited to participate as the oral source for the exercise, but did not form part of the board.

The entry test was administered to a population of 33 individuals, out of a total of 41 who had actually registered for it. Of these, one person had chosen to take the test in both LB English and French, while all of the others had chosen one language B only.

The procedure by which the test was administered was as follows: A 15-minute lecture was read aloud by the native-speaker of Spanish. The text had been previously chosen by the teacher concerned, without consulting the members of the board, on the basis of his understanding of the nature of aptitude for translation. This teacher did not actually teach translation, but had a number of year's experience in the School. Candidates listened and made notes during the course of the lecture, which lasted some fifteen minutes. They were then given the task of writing an LB summary of the content of the lecture. A time limit of 45 minutes was established, but a word limit was not specified. The candidate who had applied to study in two LBs was given extra time in order to write a summary in each of the languages. The members of the Examining board responsible for each of the languages then marked the summaries. In doing so, they were only required to make Pass/Fail decisions.

3.4.3.2 *Discussion*

As with the pre-course test described above, this examining process gave rise to many difficulties. Members of the Examining board were neither required, nor apparently did they expect to discuss or have to agree criteria for marking scripts. One member of the board offered the concept of *libertad de cátedra* as a carte blanche for total independence and autonomy. Checks of intra- or inter-rater reliability were not made, and as scores were not recorded numerically, nor did any formal written criteria exist, it was impossible to carry out even the most rudimentary post-examination analysis. In one-to-one conversations after the marking process had been completed, it was apparent that the Board members had very different philosophies as to the purpose of the test exercise, and that they did not conceive it necessary to come to any sort of consensus agreement. These differences revealed themselves in the approach taken to marking and the unbalanced pattern of passes and failures pictured in Figure 3—5.

The balance indicated by the first two columns, which are almost identical in size — 17 candidates failed and 16 passed — is very different to the picture when we consider the breakdown by languages. For English, only a minority of candidates passed (21%), whereas in French and German the figures were more than reversed, with 89% and 80% of passes, respectively. One member of the board used the mechanical error-count method criticized by Heaton (1981), while another applied a balance of this approach holistic impression marking (Heaton 1981). The third acted out of the conviction that anyone who, at the age of 25 or older wanted to start university studies, should be given every opportunity to do so. He only failed one candidate because he not demonstrated even minimal linguistic competence, and he felt he had no choice but to do so. Clearly, a great degree of inconsistency was apparent in the results.

No statistical analysis of results obtained by using this model has been published. Consequently the reliability or validity of this particular test, or of the use of this model of test in Granada, has never been demonstrated. From the relatively superficial study that we have presented here we conclude that they were highly unlikely to prove either reliable or valid due to the manner in which they were administered. This does not completely invalidate the test, though, and satisfactory administrative procedures could be designed and developed in order to satisfy criteria of reliability and validity. However, it is obvious that a number of constraints on the much-prized freedom — *libertad de cátedra* — enjoyed by board members, would have to be introduced in order to establish basic standards of inter- and intra-rater reliability.

3.5 Predicted candida te numbers

Student matriculation prior to 1992 was a further factor that had to be taken into consideration in order to begin the process of planning for the entry test. Initially, the first degree program was to adopt the same *numerus clausus* as the Diploma: 200 students in total, made up of 80, 80, and 40 for English, French and German, respectively. Despite the fact that sixteen public and private universities had at this stage begun, or announced their intention to begin to teach the degree program, a large number of applicants were still expected. It was argued within the University of Granada that the prestige held by both Granada and Barcelona would ensure high numbers of applications from all over the country at both of these centers. Moreover, the university authorities' decision that on this occasion no examination fee would be charged supported this view.

In order to set a target number of candidates, we studied data supplied by the *E.U.T.I.* office covering student applications for places on the Diploma program over the previous five academic years, from 1988-89 to 1992-93. The statistics gathered were extremely healthy, and although a figure as high as 1,108 applicants — the average for June sessions over the period studied — seemed highly unlikely, the School decided to work on a maximum of 700 possible applicants for the three languages. This decision, arguably on the high side, made it essential that the administration of the entry tests and the subsequent process of marking be as fast and as efficient as possible. Taking into account this target, and the fact that the entry test date had to fit into the timetable of university entrance examinations, and be administered on university premises meant that logistical considerations would need to play a significant part in the test design process. The question of the overall practicality of the entry test, which we will define in Chapter 7, comes to the fore when we consider the dimensions of our task.

3.6 Admission to Tran slation studies around the world

A second input to our study were those publications that dealt specifically with the question of entry to Translation studies programs around the world. However, in this field we were disappointed to encounter a near absence of publications on the subject. The two major articles dealing with the topic come from the University of Ottawa, Canada, and were apparently based on the same set of data. Bossé-Andrieu (1981) and Campagna and Dionne (1981) made detailed studies of admission to Translation studies programs in francophone countries, and most specifically in Ottawa. From their conclusions we drew a number of pointers towards possible outcomes of our aptitude test results. In particular, we found Bossé-Andrieu's conclusions with regard to the relative importance of LA and LB, and her correlation of a range of academic measures highly relevant.

3.6.1 *BOSSÉ-ANDRIEU'S STUD Y OF FRANCOPHONE COUNTRIES*

Bossé-Andrieu (1981) surveyed admission procedures to schools of translation and interpreting in Belgium, Canada, France and Switzerland. She worked within the context of centers teaching translation from and into LA French from LB English and/or

Figure 3—5 Entry test for mature students (*Actas del examen de acceso a la Universidad de aspirantes mayores de 25 años.* April 1992)

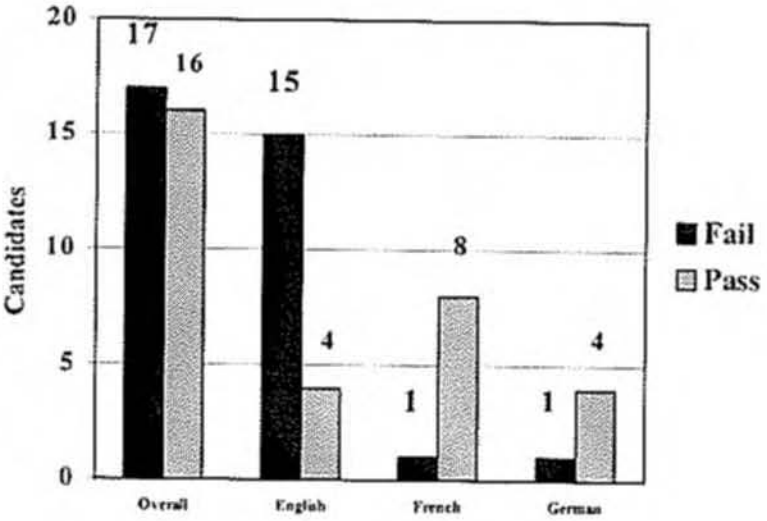
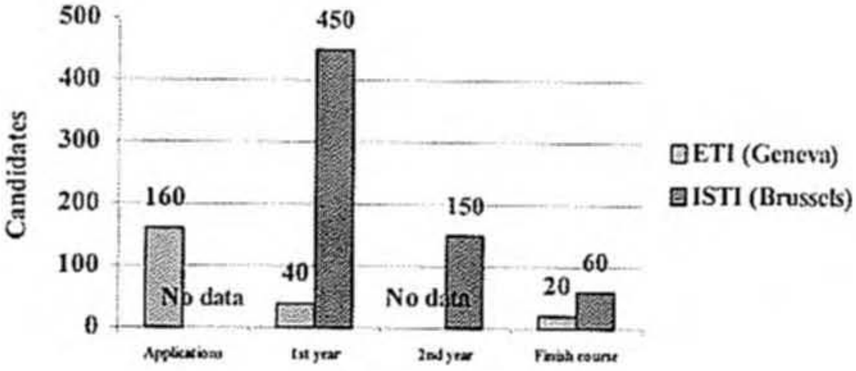


Figure 3—6 Dropout rates in two francophone countries (Bossé-Andrieu 1981:466-68)



other languages. She highlighted the wide range of entry criteria used in different centers, and the variety of procedures adopted, while focussing on the common problem of "student mortality" as depicted in the drop-out and failure rates of some of the centers. Although she is not able to offer statistical details for all of the centers in her study, we present a comparison of the two centers for which data is almost complete in Figure 3—6. No numbers are given for the total of applications in Brussels, or for the dropout rate between first and second year in Geneva.

The basic pattern can be seen to be the same in both Brussels and Geneva, and this is reflected in the dropout percentages given for the other centers. For example, the ETI at Mons experienced a 50% dropout after the second year, and among the Canadian centers Glendon College and Moncton University lost 34% and 60% of first year students, respectively; while for Ottawa, Bossé-Andrieu reported that only 40% completed their studies.

The purpose of this comparison of dropout rates and admission procedures was to contextualize a correlation study carried out at L'École de traduction et d'interprétation de l'Université d'Ottawa (ETI). This compared course results achieved in secondary school with first and second year studies in the ETI in order to shed light on the *risques d'erreur dans le recrutement des candidats* (472). It was important for her to include second year results because in Ottawa students begin Translation Studies in their second year at the university. The measures used included LA French and LB English, Translation B-A, and Secondary school achievement grades.

The study described five European and eight Canadian centers, with data drawn from published brochures and/or gathered at a conference (APFUCC 1980). Brief descriptions of entry tests were also included.

In both the European and Canadian centers selection included consideration of Secondary school marks; in some it also included first year/cycle university marks. Four out of five European centers selected on the basis of entry tests: only Brussels accepted candidates via the Belgian general University entrance test. At Mons, tests of motivation and orientation had been used previously but these were abandoned as the results were considered not to be significant. In the four remaining centers, candidates took a test in LA (French); three centers also set a test in LB (English). All four set translation tests of different linguistic combinations: B-A, at all four centers; C-A, at three; A-B, two; and A-C, one. The language tests took various forms and generally involved a synthetic approach to testing, particularly with the use of A-A summary exercises, oral or written. At one center, a more analytic approach was taken to LB testing of grammar and syntax.

Only three of the eight Canadian centers tested LA and LB; two others used non-eliminatory placement tests. A-A summary writing exercises were used by all three centers. At Ottawa these were supplemented by analytical exercises including multiple choice question (MCQ) tests in grammar, vocabulary and anglicisms. Two centers used LA error-correction exercises, and all three tested LB reading comprehension and LB grammar.

3.6.1.1 *Discussion*

Bossé-Andrieu's findings from her comparison of centers indicated a consensus on candidates' need to demonstrate their ability to write and think "well", with little agreement on what this meant in practical terms. She drew three conclusions: firstly, that — exactly like their Spanish counterparts — the centers preferred to use their own selection process in addition to the general University entrance procedures as this combination had resulted in a lowering of failure rates, although she offered no data in support of this. At the end of the 1970s, she indicated that this trend was on the increase with more and more of the Canadian centers introducing their own tests. Secondly, she found that there was little statistical evidence available on which to make reliable judgements either in support of the status of the tests being used, or against them. As in Spain, no data had been published. Thirdly, she concluded that centers placed far greater importance on candidates' skills in LA as opposed to LB, on both sides of the Atlantic. The difference being that Canadian centers were more inclined to value correctness, precision, clarity, analytical and synthetic skills, comprehension and re-expression, than their European counterparts. For them, written production took precedence, as demonstrated in documentation published by the Institut supérieur interprétariat et traduction de l'Institut catholique de Paris (ISIT), where vocabulary, *sens de la langue* and the quality of written production were the stated criteria by which candidates were judged. In a clear contrast to the prevailing situation in Spain, on both continents, for Francophone teachers of translation, LB was clearly regarded as something students could learn: LA reading and writing skills were considered a better measure of aptitude, both implicitly and explicitly.

To begin our analysis of Bossé-Andrieu's results we present the scale by which Cohen and Mannion (1989:168-69) propose that the relationships between variables expressed through correlation coefficients be interpreted (Table 3—3).

If we follow this scale, the target for a correlation coefficient in a study of this nature would be $r \geq 0.65$. However, Ehrman (1995), working in the field of aptitude testing indicated that for practical purposes she had modified the widely accepted interpretation we have been discussing. In order to take into account the highly

Table 3—3 These are crude labels that we attach to *r* values

Coefficient value r	*Description*
0.20-0.35	Of limited use.
0.35-0.65	Of little use unless corroborated by more than one study.
0.65-0.85	"...accurate enough for most purposes"
>0.85	Strong correlation

homogeneous nature of the population she was studying, Ehrman set $r \geq 0.20$ as her minimally significant correlation coefficient. From this debate, we conclude that the more detailed the information available about the demographic makeup of the cohorts studied, the more "relaxed" we can be in our interpretation of the coefficients produced.

Table 3—4 Cohort A: Students from CEGEPs (Collège d'enseignement général et professionnel) in Quebec province, and who studied first year subjects in the ETI in 1977-78 (Translated and adapted from Bossé-Andrieu 1981)

	SSch Fr	SSch Eng	SSch Overall	TR 2588-89 (French)	TRA 2311 (English)	TRA 2522 (Translation B-A)
SSch Fr	1	No data	No data	0.207	No data	0.197
SSch Eng		1	No data	No data	0.339	0.162
Ssch Overall			1	No data	No data	0.254
TR 2588-89 (French)				1	No data	0.784
TRA 2311 (English)					1	0.371
TRA 2522 (Translation B-A)						1

In the study in question, Bossé-Andrieu compared pre- and post-entry course marks in her search for potential predictors of aptitude. These had been collated from two cohorts, which we have labeled A and B. No details were presented as to the number of individuals in each group, and little was revealed of their demographic makeup. Cohort A was formed by students who had entered ETI through the Quebec secondary school system, and who had undertaken their first year of Translation Studies in the ETI. The marks analyzed in their case were: the overall secondary school score (SSch Overall), equivalent to the secondary school average (Vigneault 1998) in Spain, and the results for secondary school French (SSch Fr) and secondary school English (SSch Eng) language courses. To these were added three university course subjects which were the target measures of the study: a French language course taught within the ETI (TR 2588-89), and English language course also taught in the ETI (TR 2311), and an English into French translation course (TRA 2522). All three subjects were taught within the ETI. These three university courses would, superficially at least, appear equivalent to the current LA Spanish, LB English, and General translation B-A subjects that appear in the current degree program in Granada. We present a translation of the original data in Table 3—4, which we will now comment on in detail.

Due to the lack of demographic information available, we initially interpret the strengths of these against Cohen and Mannion's guidelines (1989:168-69). Accordingly, we see that secondary school marks for French and English, and the overall secondary school score, all of which produce positive correlations, fail to provide coefficients of more than "limited use" when correlated with marks for translation. Similarly, the correlations between secondary school French and a university French language course, and between secondary school English and a university English language course, are both below 0.35. This in itself is evidently strange. The only explanations that we can imagine are either that the academic criteria used in secondary schools and at university are very different, or that the low correlation is a factor of the difficulties individuals may have encountered in changing from the secondary to tertiary learning environments. The correlations between courses taught by ETI teachers prove to be consistently stronger than those taught by teachers from other departments. The correlation between the French language course and Translation from English into French is as high as 0.784 and that between the English language course and the same translation scores 0.371. These statistics would seem to support the first of our explanations. That is, that teachers in the university use different criteria of evaluation to those applied in secondary schools. Intuitively, this explanation does appear to have a certain face validity, especially if we consider it logical that teachers within the same school, the ETI, are also likely to share academic criteria. Homogeneity among raters, while never guaranteed is at least highly desirable.

On the other hand, we could apply Ehrman's much lower target correlation coefficient (≥0.20). We can justify this by taking into account the probability that a degree of homogeneity almost certainly existed among the members of the cohort. We then find that the two relationships that fail to achieve this standard are those between secondary school French and English language courses, and the university taught translation course (0.197 and 0.162, respectively). The low nature of these correlations may be accounted for in that translation, as an academic discipline, was presumably new to all of the students and that therefore the subject was likely to bring out different qualities in candidates. In this sense these figures would support Bossé-Andrieu in her search for an adequate aptitude test to measure candidates prior to entry to Translation studies. Our further interpretation of the results against Ehrman's criterion, can only underline the extremely high correlation between the university-taught French language and B-A translation courses (0.784).

Table 3—5 Cohort B: Students from secondary schools in Ontario who followed first year studies in the University of Ottawa in 1976-77, and first year studies in the ETI in 1977-78. (Translated and adapted from Bossé-Andrieu 1981)

	Univ Fr	Univ Eng	Univ Overall	TR 2588-89 (French)	TRA 2311 (English)	TRA 2522 (Translation B-A)
Univ Fr	1	No data	No data	0.271	No data	0.111
Univ Eng		1	No data	No data	0.305	0.603
Univ Overall			1	No data	No data	0.394
TR 2588-89 (French)				1	No data	0.964
TRA 2311 (English)					1	0.557
TRA 2522 (Trans-lation B-A)						1

If we turn now to Cohort B, we find a very similar pattern of results (Table 3—5). This cohort was composed of students who had entered university through the Ontario secondary school system, and who had studied subjects other than translation for one year within the University of Ottawa, before entering the first year of Translation studies in the ETI. The secondary school marks analyzed for Cohort A were replaced by "parallel" marks for subjects taught within the university, but outside of the ETI. These were university-taught courses in French and English language, and the overall average for first year university studies. The other three variables were the scores achieved in the same three ETI-taught subjects: French (TR 2588-89), English (TR 2311), and B-A Trans (TRA 2522).

The general pattern of the correlations is almost identical. The only significant difference lies in the much higher coefficient recorded for the relationship between the university-taught English language course and the B-A translation course. At 0.603, this almost reaches Cohen and Mannion's minimally acceptable target of ≥0.65. This figure is supported by the higher correlation achieved by this cohort for the relationship between the LB English course and B-A translation (0.557, as opposed to 0.371 for cohort A), both of which are taught within the ETI. What is of interest is the low correlation between the university-taught French language course and LA French taught within the ETI (0.271), and the even lower relationship between this and B-A translation (0.111): the lowest of all of the correlations presented, and in direct contrast to the figure of 0.964, recorded for the relationship between Language A as taught within the ETI and B-A translation. At this distance, both in time and space, we can only speculate on the underlying causes of these sharp contrasts, and it is not the purpose of this study to delve into this area. However, a clear divergence of program aims, content, and evaluation criteria would seem to have existed at the time of the study. In the light of this evidence and the experience of the entry test for mature students, which we have described earlier, a further research question, therefore, arises: **do similar differences exist between subjects taught within the University of Granada?** It is not the purpose of this study to look deeply into this issue, but we suggest that comparability of studies is an issue that requires further research.

Bossé-Andrieu's conclusion to the first part of her study was that written production skills in LA were believed to be an essential measurement of aptitude for translation. Furthermore, from the results of her correlational study, she inferred that LB was best taught in the specialized schools of translation; and that what was termed *moyenne générale* — the overall secondary school average and equivalent to the Spanish *nota media de COU y BUP* — was a better indicator of aptitude than were pre-entry course marks in LA or LB. This measure of general intelligence finds an echo in the measures used by Ehrman in her studies (1993, 1995a, 1995b) and in the pioneering work on which she based her research by Carroll (Carroll and Sapon 1959). We suggest that this latter conclusion is not clearly supported by the correlation coefficients she presents. We do so on the grounds that, in the absence of more detailed demographic information that might indicate the homogeneity of the cohorts studied, very few of her results achieve the minimally acceptable target for a correlation coefficient of ≥0.65.

3.6.1.2 *Conclusions*

Bossé-Andrieu drew her research question from the anecdotal records of colleagues and counterparts throughout the francophone world in the same way that we began our analysis of the situation in Granada. While there was little agreement on some matters, for example the wide range of instruments used in the process of test candidates prior to entry, there was an implicit understanding of aptitude for translation. Her correlation study did not describe the instruments used in measuring the course scores that it compared. These may have been single examination marks, or continuous assessment marks; they may have been integrative tests or analytic tests; they may or may not have involved exercises in translation. Nor did it offer detailed information on the size and characteristics of the individuals in each cohort. Furthermore, the relative weakness of some of the correlations gives rise to questions. A correlation as low as 0.111 between a university taught course in French language and another in Translation into French seems highly unusual, but further research into the possible nature of these correlations would seem to have been postponed *sine die*.

All this notwithstanding, Bossé-Andrieu's work represents the first, empirical analysis of measures of aptitude for translation and it has informed our study in a number of areas. Firstly, it has provided us with a counterbalance to the LB oriented concerns of teachers in Spain indicating that at least as much, if not more attention needs to be given to the testing of LA in the entry process. Secondly, it has shown the perceived importance of written expression, over and above any consideration of the so-called "passive" skill of reading comprehension. And thirdly it has thrown up a number of methodological considerations directly derived from the interpretation of a correlational study. We will return to all of these points later, when we deal with our test specification and design, and with the criteria by which we propose to evaluate our results. One further point of interest is the fact that Bossé-Andrieu has also opened up an area for further potential research, namely that of the divergence of results obtained by students studying different courses within the same university. But that must be left aside for future analysis.

3.6.2 *A PARALLEL STUDY: CAMPAGNA AND DIONNE (1981)*

At the same time as Bossé-Andrieu was carrying out her research, two of her colleagues were conducting a complementary study. Campagna and Dionne (1981) made a longitudinal study into the relationship between two measures of aptitude, an inventory of interests, and success in university studies defined as average annual levels of academic attainment and as course final academic achievement. Each study had different research objectives: while Bossé-Andrieu was primarily interested in the process of admission to programs in Translation studies with a view to minimizing student mortality, Campagna and Dionne approached the question from the position of the University Counseling Service. Their aim was to look at student mortality and academic success, by describing a possible profile of the Translation studies candidate in the hope that they would be able to detect factors indicating which candidates had an appropriate profile to ensure successful completion of the course.

The study, over the period 1977-80, produced results in line with the authors' initial hypothesis that there would be weak but discernable positive correlations between these measures. The study was based on a small population (Year 1 $N = 34$) which underwent the inevitable process of erosion (Year 2 $n= 22$; Year 3 $n = 15$) that rendered these results of no more than a descriptive value. However, their statistical analysis did indicate the overall adequacy of their hypothesis and of their research procedures.

At this point, we feel it important to state that we have been unable to ascertain whether or not the figures presented by Campagna and Dionne overlaps with those of Bossé-Andrieu. It may be that some or all of the students involved in one study were also part of the other. Attempts at correspondence with Louise Campagna, who has now retired from the University of Ottawa, have gone unanswered to date. Jacqueline Bossé-Andrieu has replied, but is unable to offer more than the data in her publication, and copies of some of the documents pertaining to the Round Table conference (APFUCC 1980; Deschamps undated) at which she gathered her initial comparative data.

Campagna and Dionne highlighted a number of points that are of interest. They coincided with Bossé-Andrieu in regarding the secondary school average as a measure of overall intellectual ability — Bossé-Andrieu had gone as far as to label this as a measure of intelligence quotient (IQ). In addition they made a specific division of aptitude into academic aptitude and verbal reasoning, and used specific tests for each. Their results supported the general predictions of the test-writers. They also highlighted the lack of prior research in the specific area.

3.6.3 *A POSTSCRIPT FROM WESTERN CANADA*

Rainey (1988) echoes the opinions of translation teachers in Spain on standards of LB proficiency attained in secondary schools, although his conclusions are more positive than those of his Spanish counterparts. He does not suggest the need for a separate entry test. In a descriptive study with no empirical content, the author looks at qualitative aspects of written production. He bases his study on quotations from the press in Britain, France Canada and US, and from writers on the subject of the language crisis. Rainey concludes that even if language crises exist neither English nor French are going to disappear. However, translation students must be aware of the abuses to which language is subject, especially for political purposes. Any publicity given to the language crisis is no bad thing if it improves awareness of, and stimulates interest in language per se. Teachers, he suggests, have to draw a fine line between "pedantic description and a laissez-faire attitude" for many language changes are "trivia or matters of personal taste".

3.6.4 *ATTITUDES IN THE US*

In the anglophone world, the experienced voice of Wilhelm Weber (1984) stands out in support of Bossé-Andrieu's findings, although his work lacks the support of empirical evidence. Weber emphasized the greater importance of the students' command of their LA.

Weber described his experience of trainee translator aptitudes. He affirmed that good grades in advanced composition and essay writing were a particularly good indication of aptitude. He put forward the belief that "writing and stylistic exercises are exercises in intellectual self-discipline and flexibility – two extremely important aptitudes in a translator (4). In addition he emphasized the importance of the command of LA. Interestingly, and without any empirical evidence to support him, he coincided with Bossé-Andrieu's results indicating the significance of secondary school results.

3.6.5 *A WORLDWIDE PERSPECTIVE*

The First *Language International* Conference (Dollerup and Loddegaard 1992) provided a wide-ranging, but unsystematic overview of entry procedures to courses in Translation studies, with contributions from countries as far afield as Austria, Israel, Uruguay, and Switzerland.

3.6.5.1 *Austria*

Mary Snell-Hornby reported (9-22) that at the Translation and Interpreting Institute Vienna, legal constraints on university entry meant the center had no legal right to select candidates once they had obtained the school-leaving certificate. She indicated that the dropout rate there was high. This placed Vienna in a situation similar to that in Spain and Belgium.

3.6.5.2 *Uruguay*

Sainz (69) reported that university entrance in Uruguay was automatic for those candidates who passed all of the corresponding subject examinations at the end of their secondary schooling. However, there was an entrance examination used to establish LB proficiency and ensure homogeneous teaching groups. This examination consisted of two short passages, for B-A (English into Spanish) and A-B (Spanish into English) translation. In addition there was a language paper in which candidates were examined in reading comprehension and grammar along the lines of the Cambridge Proficiency examination.

3.6.5.3 *Israel*

Amit-Kochavi (93-97) reported that Translation studies leading to a professional qualification were taught at postgraduate certificate level in Israel, where an entrance test was taken, which involved the translation of "two or three scientific texts".

3.6.5.4 *Switzerland: Zurich School for Translation and Interpretation*

Finally Renfer (175) described the varied entry system used in Zurich, where entry requirements covered a number of options: candidates may have successfully completed secondary schooling, or have taken certain preparatory courses, or take an entry test.

3.6.6 *JAPAN: THE QUESTION OF VOCATIONAL APTITUDE AMONG PROFESSIONAL TRANSLATORS*

In contrast to the research we have looked at so far, Szuki (1988) reports on a postal questionnaire survey into vocational aptitude aimed at distinguishing between "well-adapted" translators and "well-adapted" interpreters in Japan. The study offers a range of definitions of aptitude in the work context, all of which are based on the supposition that it is a composite of several factors:

> *... mental and physiological characteristics which are required to accomplish a job. (Ibukiyama 1969, quoted by Szuki 1988:108)*
>
> *psychological factors which contribute to the success of occupations to various degrees ... such as perception and intellect (Super 1957, quoted by Szuki 1988:108)*
>
> *(1) intellectual factors, (2) personality factors and (3) physical factors ... a compound of personal characteristics appropriate for a chosen career. (Yanai 1975, quoted by Szuki 1988:108)*

All of these Szuki qualifies in reaching a working definition:

> *I will define aptitude as including personal characteristics which help one to choose and stick with a career. (Szuki 1988:109)*

In the period June-September 1984, using a postal questionnaire survey, Szuki contacted a sample of 244 full- or part-time translators, interpreters or "dualists", i.e. people who work as both. The response rate was 38%, and among the translators this totaled 30, replies, of whom 21 were subsequently classified as "well-adapted". The criteria used to qualify individuals as such were the subjective reports of their employers and the length of their work experience, along with their personal comments on their jobs.

The questionnaire elicited information about interests and personality factors as well as basic demographic information. Using discriminant analyses, Szuki reached the conclusion that a series of factors ranked highly in the profiles of "well-adapted" translators, interpreters, and dualists, respectively.

The six most important factors in the translators were

> *(1) Have an interest in the arts, especially in writing novels, screen plays, drama, haiku and poems*
>
> *(2) Have interest in looking after others and voluntary work*
>
> *(3) Have interest in intercultural contact on the job and in daily life*
>
> *(4) Patient*
>
> *(5) Cheerful and humorous*
>
> *(6) Active (110-111)*

While the survey is very different from that which we require, we cannot ignore the fact that personality traits and interests form a part of the translator and can be studied in the search for a translator profile. This research is similar to aspects of the FL research reported by Ehrman (1995), and needs to be considered as a complement to the quantitative evaluation of skills that we are involved in.

3.6.7 *SURVEY OF FINDINGS*

Perhaps the most important difference between the situation in the francophone countries studied by Bossé-Andrieu and that in Granada is the perception held by teachers within the Diploma program of the standards of English attained by their students prior to entry. It is not the place of this study to enter into a comparison of teaching methods and standards around the world, although clearly local factors have a significant influence on developments. The fact that, for example, Canada is a bilingual country, or that most Swiss have a command of two or more languages, must influence their respective situations. However, it would seem likely that an assumption of a lower LB entry point for candidates starting to study translation in Spain could be expected if we compare them with their Canadian or Swiss counterparts. However, the same cannot be said of students in Belgium or France. Where, then, does that leave this view?

3.7 Conclusions

Our analysis of the documented and anecdotal evidence available on the practical application of entry tests for courses in translation and interpreting around the world revealed a number of major difficulties from the perspective of our study, and resolved none of them. It will be recalled that our three stated aims for this initial study were

- To identify a model of test and test specification that could be applied in the University of Granada
- To describe the underlying theoretical principles behind such a test, or those which might be applied in the preparation of a test
- To identify a research design model that might be used in the process of developing and refining an entry test.

In fact, what we conclude is that the area of entry testing for Translation studies is as yet relatively untouched in terms of empirical research, and that our hunt for a model, and a test specification, must continue in other fields. Only the matter of an appropriate research design would appear to have been partially resolved by this study: Bossé-Andrieu's correlational model would appear to be both appropriate and adequate for our purposes. Szuki (1988) has offered an interesting insight into other aspects of the issue of aptitude that could be taken further in a later research project.

In addition, our search within Translation studies has led us to a number of conclusions that will provide input to the study. These include the contrasting attitudes between teachers in Spain and elsewhere:

Table 3—6 A comparison of aspects of entry tests in Spain and abroad

Spain	*Elsewhere*
Perceived problem of poor LB (English) standards	Mixed views on relative balance of LA and LB
Synthetic model test	Variety of models
No written test specification	No published test specification
No empirical data analysis	Little empirical data analysis, uncertain quality
No research design to follow	Correlational research

In particular the debate ranged over three principal theoretical issues:

- The relative importance of the two languages involved
- The test mechanisms and marking procedures to be used, whether SYNTHETIC or ANALYTIC
- The relative importance of the active, or productive skill of written expression, against the passive, or receptive skill of reading comprehension.

The limited statistical evidence published did little to illuminate these issues, although it did shed some light on the research model that we might employ in our study.

Moreover, we have compiled a list of some of the "real-world" problems which have appeared, and which will need to be managed in the course of our study if the practicality of the testing exercise is to be ensured.

- FACE VALIDITY (teachers and applicants)
- Student numbers
- Test administration and marking logistics

All of these data lead us to continue our review of the literature into other, neighboring, fields and disciplines, although there are two important gains from this review of the context. Firstly, we have discovered a set of markers (Tables 3—4 and 3—5) that provide us with a general frame of markers on which we can model part of our research. Secondly, we have seen a test instrument in operation, the OPT, which can serve as one of the measures for the validation for our test.

4 "APTITUDE IS NOT EASY TO DEFINE"

4 "APTITUDE IS NOT EASY TO DEFINE"

4.1 A HISTORY OF APTITUDE TESTING
4.1.1 *Discussion*
4.2 AN OVERVIEW OF FLL APTITUDE
4.3 JOHN B. CARROLL: APTITUDE TESTING AND RESEARCH
4.3.1 *Distinctions between FL aptitude and aptitude for translation*
4.3.1.1 Discussion
4.3.2 *The definition of aptitude, and of related concepts*
4.3.2.1 Discussion
4.3.3 *FL aptitude: what it is, and what it is not.*
4.3.4 *The* g *factor, the* v *factor, and the* w *factor*
4.3.4.1 Discussion
4.3.5 *A research design for aptitude*
4.3.6 *Conclusions*
4.4 THE MLAT, STILL IN USE IN THE 90S
4.5 "HORSES FOR COURSES": AN EXAMPLE OF AN APTITUDE TEST DESIGNED FOR A HIGHLY SPECIFIC POPULATION
4.5.1 *Discussion*
4.6 APTITUDE FOR TRANSLATION
4.7 CONCLUSIONS

OUR REVIEW of the literature on aptitude testing in the sphere of Foreign Language Learning (FLL) had one specific objective

- ⁂ to define the concept of aptitude and the terms related to it

as well as those of continuing our search

- ⁂ to discover a model and/or specification
- ⁂ to identify underlying theoretical principles
- ⁂ to find a research design to apply in our study.

In view of the number of unresolved issues arising from our initial investigation of the immediate historical and academic contexts in which the entry test was situated we now embark on an analysis of research from fields other than Translation studies. We look at research into aptitude testing from Second and Foreign Language Learning (SLL and FLL, respectively).

We have taken the title for this chapter from Ellis (1985), a leading authority on the subject of Second Language Acquisition (SLA). And we begin by situating aptitude within the context of language learning, and the factors that influence individual learners in terms of the "route" and "rate" by, and at which they learn. Ellis lists aptitude as one of five factors recognized as distinguishing between learners: age, aptitude, cognitive style, motivation, and personality (11). In this context, he distinguishes between intelligence and aptitude by defining the former as a range of general abilities, of the kind that individual learners employ in learning all or any subjects, and the latter as a special ability. As such, he says, research has identified the influence of aptitude as a variable in language learning, and he cites Gardner (1980), although he then casts doubt on the nature of aptitude in work such as this, because we do not know which cognitive abilities constitute aptitude.

The distinction, then, between intelligence and aptitude is one that Ellis chooses to investigate further. To begin with, following Stern (1983), he suggests that aptitude is largely defined by the test used to measure it. Ellis gives the Modern Language Aptitude Test (MLAT) (Carroll and Sapon 1959) and Pimsleur's Language Aptitude Battery (PLAB) (1966) as examples of this. These batteries coincide in using tests that focus on phonetic coding ability, grammatical sensitivity, and inductive ability. Are these, therefore, to be accepted as the cognitive skills that measure FLL aptitude? Clearly, we do not know enough, but we can say that the correlations between these test batteries and other external measures of FL performance are sufficiently high as to indicate a significant degree of success (Ehrman 1993). And this has been the general pattern of research throughout the 20th century, as we will now see.

4.1 A history of aptitu de testing

(Spolsky 1995) provides FL test writers and test theoreticians with a complete history of the development of testing for language learning aptitude. In particular, he documents the earliest aptitude tests, produced during the 1920s and 30s, and describes the theoretical, practical and political aspects of fully 70 years of research and test application.

From the essential motivation "to keep prospective failures out of classes" (117), Spolsky discusses the still unanswered question as to the unitary or divisible nature of language learning ability. The debate that centers on whether or not LL ability can be divided into discretely testable abilities or skills has run, he says, throughout the history of language aptitude testing.

The panorama of teachers agonizing over dropout rates and high levels of student failure — exactly the situation we have described in the University of Granada (Chapter 3) — is the background to the first attempts to test. Spolsky quotes examples of tests from as early as 1925 (Stoddard & Vander Beke), 1928 (Luria), and 1929 (Hunt et al), all of which attempted to measure test individual abilities through sets of sub-tests.

At that early stage, Spolsky reports, the debate over the unitary or divisible nature of language ability was deemed "too complex" (Henmon et al 1929, quoted by Spolsky 1995). In further evidence on the question, he describes more fully the results of Henmon et al's work. Their findings were that low positive correlations, in the range of 0.20 to 0.60, did occur between language ability as measured by objective test, IQ, and Secondary school marks. Furthermore, they found that these Secondary school averages - drawn from all subjects studied - were better measures than either IQ or the more specific Secondary school language scores. Further studies referred to by Spolsky (Bohan, in Henmon et al 1929) found an even lower range of correlations, 0.15 to 0.50, between IQ scores on entry to modern language instruction, and later achievement grades in the languages studied. By contrast, the Princeton test found that the best correlation (0.480) described the relationship between the College Board Entrance Examinations average for French, English and Latin, and the subsequent College grades in French. This relationship was further tested by Rice (undated, quoted by Spolsky 1995:120) who taught Spanish grammar and vocabulary to 100 students and calculated the correlations between the Barry test and IQ (0.79), and with the teacher's mark (0.60); the correlation between IQ and the teacher's mark being 0.53.

Table 4—1 Summary of early language aptitude tests (adapted from Spolsky 1995)

Test	*Date*	*Components*	
Language Aptitude Test	1929	10 sub-tests (200 items	Artificial language (4); Prepositions; Memory; Affixes; Rules
Luria-Orleans Modern Language Prognosis Test	1928	Language learning trial (8 Grammar-translation lessons in French & Spanish)	Vocabulary exercises (cognates and memorization); Comprehension; vocabulary; LA English grammar; Following directions; Sound recognition; Accent knowledge
Princeton Artificial Language Test		10 vocabulary items; 6 grammar rules	Translation tasks
Barry Test		Spanish grammar & vocabulary	
Symonds Test Form A	1930		English inflection; Translation from English to Esperanto; Translation of sentences from Esperanto to English; Related words

Test	*Date*	*Components*
Symonds Test Form B	1930	Parts of speech; Translation from English to Esperanto; Translation of sentences from Esperanto to English; Artificial language
Michel & Burkhard	1934 & 1936	Memory test: Short sentences in LB German with LA English translations; Analogies test of German/English cognates; German grammar rules and examples

Symonds (1930) carried out research into the validation of the prognosis test. He concluded that the work was inherently difficult because part of the population had been eliminated after the first application of the test. He reported testing for three types of ability: general intelligence, LA, and "quick-learning" of LB. The use of translation ability as a measure of grammar knowledge was reported as being "good". A later version of this test produced correlations of 0.71 between the prognosis test and achievement tests in the foreign language. In 1933, Richardson carried out a study based on Symonds test and found that it correlated well with 1st semester Secondary school scores (0.60), and that IQ scores "added little" to the information gathered. Lau (1933) found the same level of correlation (0.60) between the Symonds test and American Council Alpha tests taken at the end of the 1st semester of Secondary school.

Michel (1934 & 1936) published further comparative work on the Symonds test and the Iowa Foreign Language Aptitude test and (with Burkhard) on a Prognosis test for German. This latter instrument was reported to give a Spearman-Brown split-half reliability score of 0.917. Their conclusions were "pessimistic" in that the Symonds test correlated well with language performance in French and Spanish, but not with German.

Maronpot made a further study of the Symonds test and correlated results with teacher's final grades in French (0.70), the scholastic average (0.51), and IQ (0.27). However, in this study the objectives of testing were to grade subjects in training, not to select or exclude them from training.

Todd applied a psychological analysis of the language learning process. The test involved a general questionnaire; a test of immediate memory span based on isolated digits and later replaced by a test of logical memory; a measure of the extent of LA vocabulary; and a test on a range of information. Later comprehension tests were added. Again, the correlations between IQ and the Secondary school average for languages was reported as being good; as was the correlation between IQ and Todd's test. The sub-tests in comprehension did not correlate as well as anticipated with school marks.

4.1.1 *DISCUSSION*

In the studies that Spolsky reports on a number of instruments and measures appear, and reappear, and the correlations between them tend to fall within similar ranges. For example, he describes Rice's work based on a correlation of scores between the Barry Test, measures of IQ, and the "teacher's mark", and presents the following results:

Other studies he describes produce similar, or poorer results, but always between three similar instruments: a test of aptitude, a measure of intelligence, and a reference to secondary schooling or University grades. The most marked level of difference between the studies is in reference to the contribution that IQ makes to the information derived: In the above example it is a good marker, although weaker than the other two. In other studies, IQ contributes little.

Table 4—2 A summary of some of the results in Rice (undated, quoted in Spolsky 1995)

	Aptitude test	*IQ*	*Teacher*
Aptitude test	1	0.79	0.60
IQ		1	0.53
Teacher			1

Clearly, our expectations in terms of the results of any correlational element to our study should not be high, but it seems reasonable to assume that even "low positive correlations" would indicate a certain degree of success. Furthermore, the nature of the instruments that we should use in the study is clear. These authors have used separate measures of IQ and school achievement, unlike Bossé-Andrieu (1981) who equated the two. In our context, we will not be able to apply a separate measure of IQ, and consequently we will need to take care in interpreting the correlations between University entrance marks and Secondary school average.

4.2 An overview of FL L aptitude

In our continuing search for a test specification, we now look at Stern's survey of FLL (1983), which seeks to describe the contribution of applied linguistics and psychological research into aptitude, to our awareness of learner differences. His analysis looks at the manner in which aptitude has been defined, and contrasts aptitude with definitions of general intelligence. In the table below, we see how Stern describes the two most significant aptitude tests: the Modern Language Aptitude Test (MLAT) (Carroll and Sapon 1959) and the Pimsleur Language Aptitude Battery (PLAB) (Pimsleur 1966).

At this point, we should say that, unfortunately, we have been unable to obtain a copy of the MLAT for our own study. This is because the test is still in commercial use. Attempts to obtain access to the MLAT via the University inter-library loan system, and from the British Library have all proved fruitless. Similarly, the University of North Carolina at Chapel Hill, where John B. Carroll is Professor of Psychology Emeritus, has been unable to assist. They do not possess a copy, for the same reasons.

Stern sets the scene for his discussion of aptitude in FLL by noting that it brought about the introduction of research techniques and standards derived principally from the field of psychology. Two of the most important researchers, Carroll and Lambert, are both psychologists and their initial involvement introduced a systematic, scientific approach to research which had previously been lacking (55).

Stern reports that the generally accepted definition of aptitude is based on the understanding that, within a wide range of learner characteristics, their exist some specific aptitudes which 'good' learners display. These enable them to achieve greater 'success' in learning the target language or languages when compared with other learners. These aptitudes are considered among the significant variables, which account for differential 'success' among learners. They are distinct from more general intelligence, in that they are specifically relevant to language learning, whether the language is the individual's mother tongue or a second or foreign language. They reveal themselves in the

Table 4—3 Constituents of second language aptitude (Stern 1983:371)

MLAT/EMLAT		PLAB	
	Ability assessed		
Test task descriptions	*Names of tests*	*Names of tests*	Test task descriptions
Learn words for numbers in an artificial language	*Number learning*	*Sound discrimination*	Learn phonetic distinctions and recognize them in different contexts
Listen to sounds and learn phonetic symbols for them	*Phonetic script*	*Sound-symbol association*	Associate sounds with written symbols
Decipher phonetically spelt English words and identify words with similar meanings	*Spelling clues*	*Rhymes*	List as many words as possible that rhyme with four given words
	The ability to discriminate, remember, interpret, and produce the phonic substance of another language. Auditory alertness. The ability to relate the phonology to forms of graphemic representation.		
Recognize the syntactic functions of words and phrases in sentences	*Words in sentences*	*Language analysis*	Make judgements with the help of translations about the meanings and rules of use of an unknown language.
	The ability to pay attention to morphological, syntactic, and semantic features of a language, to relate linguistic forms to each other, and to develop patterns, regularities, and rules from linguistic materials: linguistic (grammatical-semantic) sensitivity and an inductive learning ability.		
Learn and recall words in an artificial language	*Number learning.* *Paired associates*		

MLAT/EMLAT		PLAB	
	Ability assessed		
Test task descriptions	*Names of tests*	*Names of tests*	Test task descriptions
	Memory ability: the capacity to memorize and recall words in a new language. Rote memory. MLAT/EMLAT only. Not tapped by PLAB		
		Vocabulary	Identify the meaning of different words.
	Word knowledge, i.e., lexical competence in the first language tested in PLAB only		
		Grade-point average in academic areas	Information gathered by tester
		Interest in learning a foreign language	Short questionnaire
	PLAB contains a general school achievement and motivational component, not considered in MLAT/EMLAT as part of the concept of aptitude		

acquisition of special languages or in the learning of codes and symbol systems. Testing for the specific skills that underlie aptitude, and/or achievement, is used as a means of predicting potential success in language learning.

> *The definition of second language aptitude and its measurement depend upon underlying language teaching theories and interpretations of learner characteristics and of the language learning process. (368)*

In this context, Stern points out that both the MLAT and the PLAB are tests drawn from the audiolingual approach to FLL that was dominant in the 1950s and 60s. The skills they seek to test are clearly skills that characterize this pedagogy. Sound discrimination, sound to symbol matching, rote memory, sentence structure sensitivity, and inductive language learning capacity, are all essential elements of this approach. However, notwithstanding this criticism, the essential question that must be asked is whether the tests have any practical value. Stern's position is that they do, and that they also contribute significantly to the theoretical understanding of the concept of aptitude as a variable in the language learning process. This assertion is clearly supported by the more recent research (Ehrman 1995), which we will describe in more detail later.

Carroll and Sapon defined discrete areas of language abilities such as phonetic coding, grammatical sensitivity, rote learning, and inductive language learning. They tested these through a series of five sub-tests — number learning, phonetic script, spelling clues, words in sentences, and number learning/paired associates. Pimsleur proposed just three components — verbal intelligence, which he defined after van Els et al (1978, quoted by Stern), motivation, and auditory ability. He tested these in a six-part battery of tests — sound discrimination, sound-symbol association, rhymes, language analysis, vocabulary, grade-point average in academic areas, and interest in FLL. The extra elements added by Pimsleur are significant in that they represent the growing awareness of, and interest in non-linguistic variables in the 60s. The addition of grade-point average, which could be described as equivalent to the Spanish *media global de COU* ties in with Bossé-Andrieu's use of a similar measure in her correlational study (1981), which we discussed earlier.

As Stern says, both tests are successful, but both have been criticized by a number of researchers. Ellis (1985) rightly considers that neither of them tests the ability to communicate, being representative of the audiolingual approach.

4.3 John B. Carroll: Ap titude testing and research

John B. Carroll may well have written and researched the question of foreign language (FL) aptitude more than anyone else. His publications on the subject span almost sixty years. Recently his work has followed a natural progression away from, but building on, FL aptitude to look at the much broader field of cognitive processes in the individual (1993).

In our study we draw on some of his many publications (1974, 1978, 1981, & 1993) in which he deals with a number of aspects of aptitude important for our research:

- The definition of aptitude, and of related concepts
- The distinction between FL aptitude and aptitude for translation
- FL aptitude: what it is, and what it is not.

- The concepts of general intelligence, verbal intelligence, and word fluency, and the overlaps between these and aptitude for translation
- An empirical research design to test the validity of any measure of aptitude
- Statistical approaches generally used in aptitude research, and their implications
- Cognitive processes and aptitude
- The MLAT test battery

4.3.1 *DISTINCTIONS BETWEEN FL APTITUDE AND APTITUDE FOR TRANSLATION*

In 1978, in the context of a NATO conference on Language Interpretation and Communication, Carroll "speculated" about the "[l]inguistic abilities in translators and interpreters". He began by limiting the scope of his contribution to non-empirical research. His contact with translation and interpreting, he wrote, had mainly been as a user of conference interpreters. Elsewhere he was later to specify that the object of his research was Foreign Language aptitude, and not mother tongue language aptitude. Furthermore, he specified that his interest was strictly limited to that FL learning that takes place in more or less formal learning situations (1981). Carroll emphasized that, concerning foreign language knowledge and skills, FL aptitude is associated with rate of learning. He had taken it for granted that the two most significant constructs in FL aptitude — phonetic coding and grammatical sensitivity — would have been mastered by candidates for translation and interpreting training, despite his observations of the limited FL proficiency of some such candidates.

In this context, Carroll's first question is connected with the operational definition of translation or interpreting success:

> *What is the criterion of success, that is, how can success in translation and interpretation be measured? In the case of translation (i.e. written translation), what kinds of measures of accuracy and effectiveness could be obtained? (122)*

Without responding fully to the question he then permitted himself to enter into "speculations ... about verbal abilities and other traits that might have relevance to performance in foreign language translation and interpretation." (122) He described aspects of performance based on four situations:

- Careful written translation

- Quick, informal written translation
- Consecutive conference interpretation
- Simultaneous conference interpretation

Carroll then went on to discuss the aspects of verbal intelligence, general culture and education, and foreign language knowledge and skills that might lie behind performance in each of these. He also suggested that appropriate tests — if available — might measure these.

Carroll underscored the premise that FL aptitude, his specialty, and aptitude for translation and interpreting were two different things. Furthermore, throughout his research into FL aptitude he has eschewed any association between FL aptitude and first language acquisition and/or proficiency. However, in his remarks about general intelligence and verbal intelligence, and in subsequent research, he pointed to a number of overlaps between these areas. These suggest to us that although we may attempt to measure aptitude through a specific set of skills, we will perhaps find that such a discrete analytical approach is adequate, but insufficient, due to the complexities of the behavior we measure.

4.3.1.1 *Discussion*

Clearly, Carroll's initial position is different to that we are faced with. He is looking at the training of translators and interpreters from a range of candidates for whom linguistic competence is not an issue. Hence, his question as to the nature of the "measures of accuracy and effectiveness" that could be used as performance targets.

In the case of our study, we are dealing with potential candidate performance in the area of "careful written translation", as indicated by future examination results. This is the target criterion against which we will need to correlate scores derived from our entry test.

4.3.2 *THE DEFINITION OF APTI TUDE, AND OF RELATED CONCEPTS*

Carroll has written at some length about the "aptitude-achievement distinction" (1974). In pursuing this line, he set out to clarify the differences between two concepts that are used widely, but with little rigor. He defines aptitude as

> ... *a* concept-*a construct, if you will-referring to some constellation of* conditions, *presumably residing in the individual, that predispose him to either success or failure (or some point along the continuum between these poles) in some future activity, in particular some activity requiring new learning. (286. Original emphasis)*

Carroll (1981) qualified the debate over the distinction, or lack of distinction, between measurements of aptitude and measurements of achievement as "confused". He discussed the close correlations between these measures, which can be interpreted as meaning they measure the same abilities. However, his definitions clearly distinguish between the two concepts based on the purpose of their evaluation.

Aptitude is a measure of "basic capabilities":

> *Aptitude as a concept corresponds to the notion that in approaching a particular learning task or program, the individual may be thought of as possessing some current state of capability of learning that task-if the individual is motivated, and has the opportunity of doing so. That capability is presumed to depend on some combination of more or less enduring characteristics of the individual. (Thus,* aptitude *as conceived of here does not include motivation or interest...). (1981:84. Original emphasis)*

Moreover, he contrasts this with

> Achievement, *on the other hand, corresponds to the notion that the individual can have acquired certain specified capabilities of actual performance, these capabilities being the outcomes of the learning task or program for which the individual's aptitude may have been assessed. In the case of foreign language learning, we assume that it is possible to assess the individual's aptitude before learning is started, and then to measure the individual's degree of achievement in a foreign language after his having been exposed to the learning program for a certain period of time. (1981:84. Original emphasis)*

The essential element in this distinction lies in the recognition that "aptitude measures are always measures of *some* kind of achievement". When an individual responds to a test task some prior learning has taken place to enable them to do so. However, whatever learning that represents is different to the ability to be predicted by the aptitude test:

> *Aptitude and achievement are logically distinguishable only in the context of a specific class of achievements that are predicted by a specific type of aptitude. (1981:85)*

4.3.2.1 *Discussion*

In this context, we need to be aware of the possible range of results that may derive from our aptitude test. Whichever model of test we adopt, the element of achievement testing that will be involved in candidate performance will be a function of the similarity between the test and their FLL experience. The regulations established through the legal framework that permitted the test stated that an entry test could not judge elements of training that successful candidates would later be provided with. To this extent, we see it as inevitable that a significant element of the test will appear to the candidates to be a matter FL achievement testing. This in itself could actually be a positive factor, in that it would ensure a degree of face validity for the test.

4.3.3 *FL APTITUDE: WHAT IT IS, AND WHAT IT IS NOT.*

Carroll (1981) looks at the concept of FL aptitude and considers what it is not – i.e. it is not the same as intelligence, or even verbal intelligence, although it clearly overlaps with these factors. Furthermore, he believes that it is much more difficult a concept to measure than motivation, or interest, as it is much more difficult for the individual to access his or her awareness of the concept.

Carroll quotes Gardner & Lambert (1972:2):

> *One ... wonders about the aptitude factor if he looks back into history a bit. When everyone had to know a second language, it seems that everyone, regardless of aptitude, learned it.*

4.3.4 *THE G FACTOR, THE V FACTOR, AND THE W FACTOR*

> *I would suppose that verbal intelligence would be particularly critical in the selection of personnel for training in careful, written translation, because this type of work makes great demands on the individual's sensitivity to words and their meanings, to appropriate ways of expressing ideas, and to the nuances of verbal argumentation. (124)*

Carroll (1993) draws on the initial description of general mental ability – the so-called *g* factor – as defined by Spearman (1904), and derived from the application of IQ tests. This purported to be "...a recognized higher-order factor of cognitive abilities" but in fact, represented thirty or more "distinguishable and important mental abilities" (Carroll 1993:27).

Carroll reports that specialized abilities "are largely independent of general intelligence." (27), and that these have provoked interest in their use as predictors. He then adds to Spearman's pioneering work Thurstone's contribution in the definition and description of these mental abilities, called the "primary factors" intelligence. Finally he offers his own definition of verbal intelligence, the *v* factor, as

> *Differences in the stock of linguistic responses possessed by the individual – the wealth of the individual's past experience and training in the English language.*

In this context, Carroll defines his use of "English language" as meaning "native language".

In an earlier study (1941 reported on 1993:123), Carroll had used tests of the *v* factor that involved

- knowledge of advanced vocabulary
- sensitivity to established word usages
- sensitivity to nuances of idiomatic phrases
- ability to predict the transitional probabilities of words in phrases. This he carried out through a task type called Phrase Completion, wherein the subject was required to complete phrases like "As for . . . .", with the responses scored in terms of the frequency with which they were used by other respondents.

His remarks about the *v* factor are both revealing and disquieting from our point of view:

> *The verbal factor is one of the best and most easily established factors of intelligence. It is involved not only in tests of vocabulary knowledge, but also tests of reading comprehension, ability and facility (speed) in detecting semantic and syntactic ambiguities, and ability in writing effective, highly rated themes. It would be my assumption that effective translators and interpreters should be high in this "verbal factor". The verbal factor would be particularly demanded in written translation, where exact verbal expression in the target language would be required. For practical testing purposes, the verbal intelligence factor is best measured by a wide-range vocabulary test, with emphasis on the exact meanings of the more difficult and rarer words of a language. It may be expected that scores on such a test will be correlated with scores on tests of other aspects of verbal intelligence, such as reading comprehension tests. (123-124)*

Large overlaps exist between verbal intelligence and reading skills. Carroll does not define what he understands by "reading comprehension", but it might conceivably involve the kind of target skills that are included by Munby (1978) in his taxonomy of communicative language skills.

Carroll continues:

> *Scores on verbal intelligence tests are highly correlated with amount of education, although it cannot be assumed that a diploma from an institution of higher learning will necessarily be accompanied by high scores on verbal intelligence tests.*

The caveat is important, but in the context of an entry test, which is the last in a long and intense period of examination such as our candidates undergo during the months leading up to taking *selectividad*, we must ask whether we are really testing for a specific discrete aptitude or whether the University entrance examination, or the Secondary school average, are satisfactory measures in themselves.

General culture and education are closely related to verbal intelligence. Appropriate tests were not available at the time of writing his article, and in view of this, Carroll proposed that "[e]ven if the tests assess nothing more than the level of factual information that the individual has acquired, they are likely to be predictive of the success of a translator/interpreter in meeting the demands of a wide variety of contexts…"

Success, however, might be evaluated according to any one or more of several kinds of criteria. Typically, it would be evaluated in terms of adequacy of performance relative to an ideal standard — i.e. criterion-referenced — or relative to the performance of others attempting the same activity — i.e. norm-referenced. (286)

Clark Hull's *Aptitude testing* (1928:67)

> *... aptitude tests may be divided into (1) those which attempt to duplicate all in one test the essential activities of the occupation, and (2) those which are designed to isolate and measure separately the component traits supposed to constitute the determiners of success in the aptitude in question. (294-295)*

Carroll then goes on to examine the MLAT against these criteria and finds that "[t]wo of the tests, however, require little or no learning during the actual taking of the tests, and are addressed to the measurement of certain developed abilities." Specifically, Test 4, Words in Sentences, is a test of ability to solve grammatical analogies, that is, given two sentences with a certain grammatical element identified in one, to find an element with a similar grammatical function in the other. He states that "[this test] … functions as an excellent predictor of success in foreign language learning." (295).

Carroll then goes on to speculate about the application of Thurstone's "*w*" factor of productive verbal abilities. "Word fluency", measured by various tests designed to explore the individual's ability to manipulate orthographic materials has given rise to the definitions of three fluencies:

✲ Ideational fluency, defined as "the facility to call up ideas wherein quantity and quality of idea is emphasized".

- Expressional fluency, defined as "the ability to think rapidly of appropriate wording for ideas".
- Associational fluency, defined as "the ability to produce words from a restricted area of meaning". (French, Ekstrom & Price 1963)

Carroll mentions that these fluencies tie in with Cattell's work into the "*g*" factor (1971) and have been linked by Cattell to personality traits.

Carroll quotes Ekstrom (1973:25):

> *There appears to be no evidence to suggest that this factor [associational fluency] is confined to the English language. It would be interesting to determine if the ability to produce an appropriate word when translating a well-known foreign language would involve associational fluency.*

Carroll links this to work in the field of cognitive abilities, in particular to information processing (Carroll 1967a) and to speech shadowing (Cherry 1953, quoted in Carroll 1973).

4.3.4.1 *Discussion*

What, then, are the elements that we can derive from Carroll's discussion of these three factors, which can serve us as a means of specifying the elements for inclusion in our test? The manner in which Carroll describes these factors leads us to believe that the skills specified by Munby (1978) could serve as a basic taxonomy of operationally defined language abilities representative of the *v* and *w* factors, both of which could and should be incorporated into our entry test of aptitude for translation.

4.3.5 *A RESEARCH DESIGN FOR APTITUDE*

Building on the aptitude-achievement distinction we have mentioned earlier, itself a crucial part of the design, Carroll (1981) presents a series of conditions by which an aptitude test can be judged. These amount to an empirical research model for aptitude testing.

> *... if aptitude for a learning task is measured prior to an individual's engaging in that task, and if achievement on the task is measured after a given amount of exposure to the learning task, the concepts of aptitude and achievement are operationally distinguishable. The validity of the aptitude measurement is judged in terms of the degree to which it predicts the final measure of achievement. (287)*

Consequently, the distinction is conceptual, rather than instrumental. Aptitude is measured before learning, and should remain constant over the period of learning; achievement could be expected to measure zero before learning, and would be expected to increase over the period of learning. The precise instruments used to measure aptitude or achievement do not enter this part of the research design. A so-called 'achievement' test could conceivably be a measure of aptitude; so could

> *... data on prior performance in activities similar to those for which we wish to predict success, and information derived from procedures for assessing personality, interest, attitude, physical prowess, psychological state, etc. ... interest and prior school achievement." (287)*

The six conditions are that

- The mean and variance of A_{ix} (aptitude at the outset of instruction) have values indicating reliable true-score variance, that is, the group is heterogeneous in aptitude.

- The mean and variance of B_{ix} (achievement at the outset of instruction) have values indicating that achievement is zero or essentially chance, and any variance is random error.

- The correlation between A_{ix} and B_{ix} is not significantly different from zero (because the variance of B_{ix} represents only random error).

- The mean and variance of A_{fx} are not significantly different from the corresponding values for A_{ix}, that is, there has been no significant change in the average and spread of aptitude of the group from time t_i to time t_f. Furthermore, the correlation between A_{ix} and A_{fx} is not significantly different from unity when corrected for attenuation; i.e. aptitude is essentially constant for the individual.

- The mean of B_{fx} is significantly different from the mean of B_{ix} in a desired direction, that is, on the average there has been at least some learning of task x by members of the group. Furthermore, the variance of B_{fx} is such as to indicate reliable differences in achievement.

- The correlations between A_{ix} and B_{fx} is significantly different from zero; that is measurement on A_{ix} are significant predictors of scores in B_{fx}. (288)

These conditions Carroll represents graphically as in the figure below.

Table 4—4 Carroll's six conditions to test the validity of a measure of aptitude (1981)

Initial time	*Period of learning on task*	*Final time*
t_i	x	t_f
Aptitude test A_{ix}		Aptitude test A_{fx}
(Condition 1: Reliable true score variance in A_{ix})	(Condition 4: No change from A_{ix} to A_{fx})	
(Condition 3: $r_{AixBix} = 0$)	(Condition 6: $r^2_{AixBfx} > 0$)	
Achievement test B_{ix}		Achievement test B_{fx}
(Condition 2: No reliable true score variance in B_{ix})	(Condition 5: Significant change in mean from B_{ix} to B_{fx})	

Carroll recognizes that all six conditions are unlikely ever to be met in a specific research project. Indeed, it may be impossible to meet them as aptitude research is normally carried out in "real-world" contexts, this being one of its principle failings. The reasons that conditions cannot be met are often to do with face validity. To begin with, the application of an achievement test at time t, would be unusual to say the least, and would lack face validity from the candidates' point of view. Accordingly, we assume that condition 2 holds, without actually attempting to measure it. Similarly, we do not normally administer an aptitude test at time $t_{f,}$ but simply assume that condition 4 holds. What is more, the validity of condition 3, whether it is demonstrated or not, does not mean that test A_{ix} is not a predictor of Achievement B_{fx}.

The key to the value of any aptitude test is in demonstrating that conditions 1, 5, and 6 hold. However, again, the difficulty may lie in fully testing these. If condition 2 cannot be met because no test B_{ix} has been administered, then condition 5 cannot be proven either. We are left, then, with conditions 1 and 6 as the foundations on which to base any decision as to the validity of an aptitude test.

These six "ideal conditions" (289) are further complicated, as Carroll describes, through the ways and means by which aptitude tests are designed:

> *... a shotgun approach is the method of choice: 'try anything and see what works' seems to be the watchword." (293)*

However, from his exhaustive study of aptitude tests, he finds that, procedurally,

> *... one must make an analysis of the components of the activity: what sensory and perceptual sensitivities, motor coordinations, learnings, dispositions, attitudes, etc., are prerequisites for success in the activity, or at least, if possessed, make success more likely? (293)*

Reporting on the development of the MLAT, he indicates that he had put this procedure into practice:

> *The primary consideration in selecting and devising the tests of the initial [MLAT] trial battery was to include a variety of tests each of which promised to measure some aspect of the complex traits deemed requisite for success in the criterion performance. (293)*

Both in the test and in the learning situation there might be common "elements" such as an ability to hold in memory certain types of associations between sounds and meanings, to notice certain grammatical attributes of a sentence, or to identify and recognize certain sounds. In other words, the tasks chosen for the aptitude tests were those which were regarded as having *process structures* similar to, or even identical with, the process structures exemplified in actual learning tasks (1974:294). Therefore, the tests measured the individual's ability to perform the psychological functions embedded in the criterion learning tasks. The theoretical basis for assuming similarities between aptitude tasks and criterion tasks might be of the vaguest intuitive sort; what mattered was the empirical confirmation of one's intuitions by standard test validation procedures.

4.3.6 *CONCLUSIONS*

The theoretical concept of aptitude is one that has given rise to, and continues to give rise to enormous debate. From our reading of much of Carroll's lifelong investigation of the concept we have reached a number of conclusions that we feel are satisfactory theoretical premises on which to base our study. Furthermore, we believe we have encountered a suitable research design model, which we can employ.

The divisibility of aptitude has still to be proved in empirical terms, but it would seem that there is a sound base for assuming that a discrete approach to skills that combine to make up a global aptitude is valid. The overlaps between such skills and the wider ranging concept of general intelligence, the *g* factor, is accepted, and while it is beyond the scope of this study to enter into the dimension and nature of these overlaps, they cannot be ignored in our judgements drawn from test results. If we adopt Carroll's approach, we can say that the choice of a discrete sub-set of skills might, intuitively, form a part of the process structures involved in our criterion task, namely General Translation A-B and B-A. The exact choice of the skills needs to be documented, but will probably remain unproven until such time as we are able to identify the underlying cognitive abilities and processes empirically. However, the aim of our research is to validate the use of a test instrument, which in itself is a part of the verification of our more or less intuitive choice.

In the studies that Spolsky reviews, we find that measures of IQ, and of secondary school achievement tend to achieve low positive correlations with tests of FL aptitude.

In the final parts of this chapter we will return to the MLAT, and look at the conclusions drawn from research based on a model similar to that we have described above, and in which a number of practical conclusions of relevance to our study appear. We will then describe, with a degree of anonymity, an aptitude test used in a specific context for specific purposes, in order to demonstrate the reality of such tests.

4.4 The MLAT, still in use in the 90s

The durability of the Modern Language Aptitude Test (Carroll & Sapon 1959) — still in the mid 90's a standard general measure of language learning aptitude at the US Foreign Service Institute (FSI) — has been amply demonstrated by Madeline Ehrman (1993, 1995a, 1995b). Recently published articles present different aspects of a large-scale, longitudinal study (n = 1000). Ehrman (1995a:2) describes, both in quantitative and qualitative terms, the results of a range of measures, including the Modern Language Aptitude Test (MLAT) when tested on a population of adult, native-speaker (NS) English learners of a range of foreign languages.

Given the doubts expressed by some colleagues at the FSI, Ehrman (1995a) questioned the value of the MLAT as a measure of potential student success in the context of a language learning environment, which had left behind the audio-lingual methodology that was prevalent at the time of its launch. Her results showed that the MLAT continued to be the best predictor of end-of-training success with correlations between the INDEX SCORE ranging from 0.34 to 0.55, according to the target LANGUAGE CATEGORY being analyzed and the SKILL AREA being tested.

In further studies based on the same sample, Ehrman covers a lot of ground in the correlation of measures of cognitive aptitude with other language learning variables. One objective, similar to that of Campagna and Dionne (1981) was to provide measures which offered improved student counseling. Secondly, she hoped to enable greater student autonomy in the learning process, and to establish the attributes of successful language training at higher levels of proficiency. Her work dealt with the questions of personality and language learning (1993) and with the affective and motivational variables which she described as "Cognition Plus" (1995b). Using this large, but as she recognizes somewhat atypical sample drawn from the FSI and other American government agencies, she describes correlations between the MLAT and a set of measures adapted for her study or used according to their original specifications. These were: the Affective Survey (Ehrman and Oxford 1991); the Hartmann Boundary Questionnaire (Hartmann 1991); The National Association of Secondary Schools Principals' Learning Style Profile (Keefe & Monk, with Letteri, Languis & Dunn 1981); the Myers-Briggs Type Indicator (Myers & Mccaulley 1985); the Type Differentiation Indicator (Saunders 1989); and the Strategy Inventory for Language Learning (Oxford 1989a). These course initial measures were correlated with two measures of end-of-training proficiency one in speaking and the other in reading.

Ehrman (1995) recognizes the narrow range of subjects in her sample and in doing so makes a significant methodological decision. In total, there were 855 subjects, with an average age of 39 (SD = 9), of whom 99% were LA English speakers. Most of the subjects were classified as "highly educated": 40% held masters degrees, a further 35% bachelors degrees, 9% had doctorates or law degrees. Inevitably the sameness of the sample makes generalizations more difficult to support. Nonetheless, certain findings are of interest. Firstly, the most significant correlations between the variables measured and end-of-training performance were with the MLAT index score and the MLAT subsets. While the affective and personality variables did correlate with performance, these correlations were of a lower order. Correlations considered worthy of inclusion in her table of results were those => 0.20. She proposed such a low level of correlation be accepted as a valid description on the grounds that her study was carried out with a homogeneous group, and that this in itself would lead to lower correlations than might be derived from a more heterogeneous sample.

Given this proviso, the very strength of these correlations leads Ehrman to revise the widely held view that the MLAT has lost its value as a measure of aptitude with the changes that have taken place in language teaching methodology over the 1970s and 1980s. It would appear from these results that the MLAT is, at the very least, measuring elements of cognition, which are not limited in their application to the audio-lingual method, for which the test was designed.

The significance of this study for our research lies in two directions. Firstly, comes the methodological acceptance of relatively weak correlations as being of value in a study based on a homogenous, in this case "highly educated", sample. In our study, we also deal with an apparently homogenous sample of relatively highly educated learners. What is more, the degree of homogeneity among the individuals in our sample can be quantified, in that the data on their university entrance examination scores and on their secondary school averages are available to us. Secondly, it raises the question of the general intelligence component present in the MLAT and, possibly its effect on our measure of Aptitude. The existence of such a high "g" factor in the linguistically based test would suggest that a measure of general intelligence might be a significant component of, or even satisfactory alternative to, a specific measure of Aptitude.

This hypothesis supports the earlier results of Bossé-Andrieu (1981) and Campagna & Dionne (1981).

4.5 "Horses for course s": an example of an aptitude test designed for a highly specific population

It is difficult to obtain free access to examples of commercially produced aptitude tests as their widespread publication would clearly invalidate them. We have only been able to learn at second-hand about one test and, in order to do so, we gave assurances that we would not attempt to reproduce any of the tasks included in the test. The document in question is not commercially available, but is an in-house test used by a government agency. For the purposes of this study, we will describe the context in which this test is used, and outline the test components. We will call this test the Army Aptitude Test (AAT) as it was designed to predict the FL performance of army personnel on intensive language courses. The version we describe dates from the mid 70s.

The AAT comprises four sub-tests: (A) Biodata, (B) Identification of word stress, (C) FL grammar, and (D) FL translation. All four parts use objectively marked, 4-foil multiple choice (MCQ) items. Parts (A) and (D) are based on written stimuli, while (B) and (C) are both listening tests. The rubric for each section is given in the question booklet. In general, these instructions are wordy, and they use much of what might be termed jargon, in the form of grammatical terminology. We suggest that this presupposes a certain level and type of education on the part of the candidates. In most cases, examples are given, and candidates are able to work through them, and correct them if they make a mistake before they complete the task. Responses are all made on an answer sheet. The time limits for each part are clearly set out, either on the tape or in the test booklet, and the total number of items on the test is well over 100.

The rubric for (A) Biodata explains that this sub-test is designed to gather personal information and an impression of the candidate's attitude to FL learning. There are seven items, and the first asks for the candidate's impression of the level of difficulty entailed in FLL, and the second attempts to ascertain the candidate's interest in FLL as opposed to other areas of studies. Items three, four, and five ask about previous experience of studying their first language, English, Mathematics, and their general level of education. Presumably, the test writer believes that some sort of correlation may exist between these measures and FL aptitude. The two remaining items focus on study skills and previous FLL experience.

The first of the two listening sub-tests is based on identifying the "odd-man-out" when comparing four FL words of a similar phonetic composition, in which the word stress differs for one of the four. There are 18 of these items, and each is heard once only.

Invented words and phrases in a language similar to English, but with clear differences, form the basis of the FL grammar listening sub-test which, in its turn, is divided into four parts. In three of these rules of grammar are presented and then tested, the fourth test is a synthesis of the previous three. The rules deal with aspects of syntax and morphology. Before beginning part four, candidates read a summary of the rules, and are instructed to study them for three minutes before they answer the questions. Again, candidates hear each item once only.

The final sub-test is based on visual stimuli with four-foil MCQ items. Candidates have a worked example in English to start with, and then move on to look at phrases and sentences in the invented language, which they are by now familiar with. Their task is to identify the appropriate sentence form to describe a single scene visual. The alternatives given offer misapplications of the grammar rules they have previously learned.

4.5.1 *DISCUSSION*

As we mentioned earlier, Stern (1983) suggests that tests of FL aptitude indicate underlying assumptions about FL processes. In this case, among these assumptions we find a number that have to do with the nature of the candidate population that the test was designed for — particularly in terms of educational background and familiarity with grammatical terminology — and others linked to the audiolingual FL methodology still dominant at that time. While the test itself is not applicable to our situation, it does have a number of features that we can draw on in the design of our test.

4.6 Aptitude for trans lation

Little or no research has specifically been dedicated to aptitude for translation. As we saw earlier, the most significant quantitative work on the subject was carried out in the late seventies at the École de traducteurs et d'interprètes of the University of Ottawa, Canada. Bossé-Andrieu (1981) presented data gathered in a survey of four francophone countries and offered a detailed description of the correlations obtained by comparing a range of measures of language performance and "general intelligence" drawn from university and secondary school studies. Her colleagues Campagna & Dionne (1981), apparently drawing on data from the same sample, studied the question of aptitude from the point of view of the University Counseling service. More recent work is either of a less rigorous nature, such as the information appearing in Dollerup & Loddegaard (1992), or is as yet unpublished (Reported in *Language Testing Update* 14; Szuki 1988; Fox 1997a; 1997b). By contrast, the most recent quantitative work in the field of aptitude is thoroughly documented: Ehrman (1993, 1995a, 1995b) offers much which in both theoretical and methodological terms is of value to our study.

However, we can state that in our reading of research into aptitude in the field of Foreign language learning, we have encountered a range of theoretical premises about aptitude in general which we can apply to the study of aptitude for translation. We also have a clear research design on which to model our study.

Our next objective is to define a series of operational skills that we can use in our test in order to establish the level of aptitude candidates are able to demonstrate, and we believe that these, in line with Carroll's premises with regard to the MLAT, must be found in an analysis of the process structures underling translation.

4.7 Conclusions

In this chapter, we have found that the prevailing view of aptitude is one based on a divisible concept, in which different types of aptitude exist, and are seen as overlapping. While it is considered possible to measure aptitude, it is clear that the results we can expect will not necessarily be as clear-cut as we might want them to be. Our search now moves on the a study of the translation process.

5 THE TRANSLATOR "A S EXPERT READER"

5 THE TRANSLATOR "AS EXPERT READER"

5.1 TRANSLATION AND LITERATURE

5.1.1 *Literary translation and the translation product*

5.1.2 *Conclusions*

5.2 THE READING PROCESS

5.3 FACTORS THAT INFLUENCE THE TRANSLATION PROCESS

5.4 QUALITATIVE RESEARCH INTO THE TRANSLATION PROCESS

5.4.1 *Discussion*

5.5 A PROCESS-ORIENTED APPROACH TO TRANSLATOR TRAINING

5.5.1 *A sequential model of translation*

5.5.1.1 Discussion

5.5.2 *The translator "as expert reader"*

5.5.2.1 Discussion

5.6 CONCLUSIONS

THE SPECIFIC objective of our research into the literature on translation, the translation process, and the didactics of translation is

- ✲ to identify a set of operationally defined skills and/or sub-skills, which we can employ in the specification of our model entry test.

We are looking for a series of skills that represent part of the process structure underlying translation, and which are susceptible to measurement in a test instrument.

In Chapter 3, we described the immediate historical context in which the degree program in Translation and Interpreting was established. There, we drew as accurate a picture as possible of the circumstances in which, both in Spain and abroad, certain concepts of the nature of aptitude for translation were put into practice through the admission procedures applied by different centers. Unfortunately, in our survey we failed to identify even one empirically proven test specification, and we found seriously flaws in the working of the only model test we were able to investigate. However, in our readings in the field of FL aptitude we encountered a theoretical base on which to structure a test and a research design model which we can apply.

Our next step, therefore, is to review research into translation theory and into the practice of training translators. Here we are searching for evidence of the underlying assumptions as to the operational skills demonstrated by translators in order to be able to produce our test specification.

We begin this chapter by looking at the popular, but in our opinion crude overview of translation as seen from the perspective of literary translation. Gile (1995:23) describes this approach as the classical model of translation in which a linear progression describes the relationship between author, translator, and reader. Essentially this is the perspective of many translators of literary texts, who appear to believe that translators are born, and not made, and whose essential interest is in the translation as a product.

We then move on to describe the opposite position through the contributions of a number of significant practitioners and researchers in the field (Lörscher 1986, 1991, and 1992; Bell 1991; Reiss 1992). All three place an emphasis on translation as a process, and their work provides us with insights into that process, that matches our research design model (Chapter 4). To this end, we adopt the definitions made by Bell, in which he distinguishes between three meanings of the word "translation":

- ✲ *translating: the process (to translate; the activity rather than the tangible object);*

❊ *a translation: the product of the process of translating (i.e. the translated text);*

❊ *translation: the abstract concept which encompasses both the process of translating and the product of that process. (1991:12-14)*

Finally we look at the work of two specialists in the didactics of translation Gile (1995); and Beeby Lonsdale (1996). Both, implicitly or explicitly, follow the process-oriented approach. From their work, we attempt to draw conclusions as to the manner in which we should view the question of aptitude. In particular we seek out the distinctions both writers make between an analytic and a synthetic view of translation.

5.1 Translation and lit erature

As we have mentioned, there are two opposing schools of thought as regards translation. One of these believes that translators are born and not made According to Eugene A. Nida, this product-oriented position is taken by *Translation Review*, the journal of the American Literary Translators' Association (Interviewed by Robinson 1995:109). Essentially, this approach takes a synthetic view of translation, which would describe aptitude for translation as a unitary concept: "you either have it, or you don't". An alternative point of view, is that there is a case for describing the acquisition of translation skills as a progression, or learning cycle, such as that described in the Table 5—1. This is the theoretical base from which we begin This approach is analytical and views aptitude for translation as divisible, and therefore susceptible to teaching and learning.

Table 5—1 The four stage learning process (Dilts 1994:80)

Stage 1	Unconscious incompetence
Stage 2	Conscious incompetence
Stage 3	Conscious competence
Stage 4	Unconscious competence

Dilts compares the progression seen here with that of learning to drive a car, or to type. At the outset, the individual is not aware of what is involved; they only know that they cannot "do it". Progressing to stage 2, means an important advance, in that the awareness that one is incompetent as a translator is the first step on the road to achieving professional development. It is perhaps the most important stage for the teacher of translation, for it represents the period, in which the translator has, and can learn the most. Instruction in individual skills can begin at this stage.

"Conscious competence" is probably the most realistic target for translation teachers to aim towards, within the limitations of the academic context. If the majority of students of translation are able to achieve a degree of conscious competence in their application of translation skills, they should then be able to make their way in the non-academic professional environment. Only with time and practice will they later achieve the level of "unconscious competence" (Stage 4) that goes with being a practicing professional translator.

The readers and writers of the *Translation Review* may be fortunate enough to have been able to accede to the fourth stage without having to pass through the other three. For the rest of us, and in this instance we believe that means almost all of those who choose to pass through training institutions of one sort or another - the unconsciously competent do not conceive of the need for training - for the rest of us the progression involves developing awareness to move us from Stage 1 to Stage 2. This is followed by theoretical and practical learning to move us on from Stage 2 to Stage 3, and finally, by unlimited practice to enable us to reach Stage 4. Stages 1, 2 and 3 are all within the scope of learning institutions, and all imply an analytical approach to the translating process. Stage 4 is an individual achievement that can only really be reached in the professional world, although the volume of controlled practice that can be generated within the teaching context is clearly of importance.

5.1.1 *LITERARY TRANSLATION AND THE TRANSLATION PRODUCT*

What, then, do translators of literary texts believe to be essential requirements to carry out their task? West (1992) and Muir (1992) offer some interesting insights into the question. The former summarizes his opinions in a typically English metaphor that reveals more about its author than about translation:

> *But then translation is the art of the impossible. Practitioners hope not for a bull, but for a decent inner ... (West 1992)*

A metaphor drawn from a pub game such as darts, seems to deposit the author's approach to translation in the category of leisure activities. While this may not impede publication of translations, it certainly does little to advance translation studies.

In a series of curiously chosen verbs, West goes on to reveal much of his attitude towards the translator and the translation process:

> ...*the translator has* to catch the drama... [emphasis added]

Ten lines later we find:

> *the translator has* to catch the note *of mingled love, concern and reassurance*... [emphasis added]

And four lines further on:

> *And always the translator has to* keep an eagle eye on *the detail:*... [emphasis added].

These verbs show little of the translation strategies writers such as Vázquez-Ayora (1977) would propose, demonstrating a more intuitive approach. In the same article, when he goes on to compare three published versions of the same passage, West does so with little comment on the variety that the translators demonstrate:

> *According to Jackson Knight he has dislodged the foundations and is now shattering the walls. In Day Lewis [...] he is goring and tossing them. Fitzgerald is shaking them from their beds,...*
>
> *So our ideal translator has to honor the drama, the characterization and the details, detail of course including every important poetic resource he can see and can do something about. [...] the translator of Virgil has often to concede defeat before he starts.*

The note of resignation might be that which would lead West to conclude, with many others, that translators are born, and not trained, and therefore that translator training is a futile occupation.

In the same collection Muir (1992) discusses the translation of Calderón's tragedies. He states that translating Calderón is a task which "demand[s] other skills" and draws on "what one learns from rehearsal and performance". "Our impossible task", he adds involves a wealth of knowledge , the lack of which clearly diminishes the value of the target text. He cites Jill Booty's translation of Lope de Vega from Spanish verse into English prose as an example: "No one would suspect ... that Lope was a great poet".

From these two eminent academics, we conclude that the translator must have a breadth of knowledge far beyond the purely linguistic, and that they can only acquire that knowledge with time and maturity. These are two propositions that few would argue with — indeed they are clearly encompassed by Gile when he describes extralinguistic knowledge as one of the key components of the "comprehension equation" (1995:79). However, even if we accept that these are unattainable — does Stage 4 actually exist? — that does not mean we should stop training translators.

5.1.2 *CONCLUSIONS*

The attitudes towards translation evident in these few examples taken from the writings of literary translators reveal nothing of use to our study. The translation is described as a product, and the process that goes into its creation is ignored or glossed. Aptitude for translation is apparently a unitary concept, and therefore not accessible for analysis through the type of test instrument we wish to employ. There is little that we can find of value here, and so our search must move on, initially into the realm of literary theory.

5.2 The reading proce ss

Research in the field of literary theory provides Translation studies with a firm grounding in the approach derived from Applied linguistics and from Speech Act theory. The theoretical writings of Booth (1961), Iser (1974), and put into practice by others (Traugott and Pratt 1980; Robinson 1981) reveal this. All of these authors, and many others, take the position that reading is an active process, in which the reader must participate. This means that the classical terms of active and passive, as applied to the four primary language skills, writing and speaking, listening and reading respectively no longer apply. Perhaps terms that are more appropriate are productive and receptive, which at least move us further towards accurate descriptions.

In the research we have mentioned, the complimentary concepts of "implied author" (Booth 1961:73) and "implied reader" (Iser 1974:xiii), offer the essential framework of the communicative relationship existing through the medium of a written text which the translator must work with. This relationship involves a series of figures:

- the author in person
- the implied author, a persona perceived by the reader who may or may not be a character in the text, and who may, or may not identify in some way or other with the author in person.
- the implied reader, i.e. the profile of a reader to whom the author consciously or unconsciously addresses the text
- the reader in person, the individual reading the text.

The fullest possible communication between the writer and the reader of a text takes place when the reader is able to identify with the profile, or even the role of the implied reader. It is in this context that the decoding process which reading is, most completely fulfils the expectations of the author. The reading process must be an active one, and the reader needs to participate in negotiating a role and a meaning through the text.

When we add the factor of translation to this complex, participative process, we clearly complicate the relationship considerably. The translator, as individual, has to assume a number of roles that include those in the list we have just presented, and others. To begin with, we have

- the translator as individual; the translator as receptor of the text who must fulfil the role of implied reader of the source text.

If the translator is not able to identify as closely as possible with the implied reader, their reading will be less than complete. Consequently, their comprehension of the text will be inadequate when they have to fulfil the role of

- implied author of the TT.

We distinguish this role from that of implied author of the ST because the context in which it is carried out is clearly different.

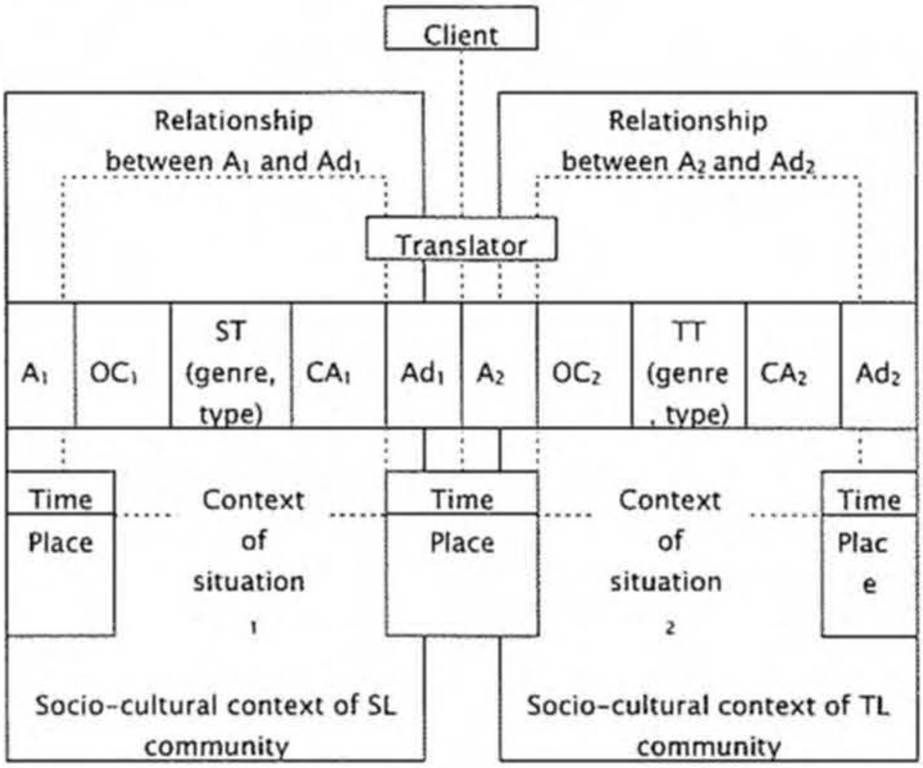
5.3 Factors that influe nce the translation process

Obviously, this relationship is complicated by the alterations of language and culture which translation of any text implies. Reiss (1992) has more fully developed this "complication" in her work on the translation process, shown in Figure 5—1.

Here she applies a theory of language usage as the basis for a theory of translation, and proposes a model that draws together those factors that influence the translation process. In the first instance, Reiss distinguishes between the sociocultural context of the Source language (SL) community and that of the Target language (TL) community. She situates the client at the junction of these two, as the instigator of the need to communicate across this boundary. A parallel relationship exists, she suggests, between the situational contexts in which source text (ST) and target text (TT) serve as vehicles of communication between author(s) and reader(s), which can be defined in terms of time and place. This parallel relationship forms a part of the boundary between sociocultural contexts in that it is also of importance at the time when the translation process takes place as it is a function of the actions of the client.

Within the strict limits of the relationships that exist within the translation process itself, Reiss establishes a chain of what she calls interactive communicative offers and acts. To begin with, she recognizes the relationships existing between the ST author and ST reader, which we equate with the figures of the implied

Figure 5—1 A model of factors that influence translation (translated and adapted from Reiss 1992)



author and implied reader, and that existing between the TT text author and TT reader. In this schema, the translator fulfils both the role of ST reader and TT author. Reiss, giving it a degree of rich detail further analyzes the nature of the author-reader relationships, which aid our understanding of the translation decision-making process.

The ST author (A_1) makes what she terms an "offer of communication" (OC_1) which is realized in the form of a SL text belonging to a certain genre or type (ST). This text performs an act of communication (CA_1) directed towards the ST reader (Ad_1), who responds. The nature of the response will be a function of the author's success in matching the offer of communication with the act, and of tailoring the text to provoke the desired reaction in the reader.

Exactly the same process is carried out on the other side of the sociocultural boundary in terms of the relationship between the author and reader of the TT. However, there is one significant difference in that the offer of communication (OC_2) may well be very different to that of the ST. The client may well frame this offer of communication for the translator, or the decision may remain ambiguous resulting in the translator having the final say. Whichever is the case, the translator must make decisions. These have to do with the function of the text within each of the acts of communication, that in the SL (CA_1) and that in the TL(CA_2); they must choose the nature of the translation process they are going to undertake; and they must opt for a SL or a TL "leaning" in their translation process and product.

When we consider the complexities of this model of the translation process, we uncover the significance of the translator's role. The translator is both reader and author. The translator is reader of the source text - and as such to achieve the most accurate reading of the text possible they must interpret the role of implied reader as fully as possible. Moreover, the translator is the author of the target text - for which they assume the mantle of implied author, for an implied readership that is most certainly going to be different to that of the source text.

When the linguistic, sociocultural, and situational differences are minimal, as between languages of the same family, in similar sociocultural, historical and geographical contexts, the demands on the translator will nonetheless be significant. We see this in terms of the compensations that may be needed to render the source text acceptable to its target readership. When these differences are greater, the demands on the translator are magnified beyond proportion. The essence of any translator's capacity to convey the author's intentions lies in the first instance in the sophisticated reading of the source text, hence our choice of reading skills as a measure of aptitude for translation. Reading is clearly crucial in the process structure that underlies translation.

Theorists have developed a range of approaches to mental processes involved in translation. Bell (1991) adopts the most powerful metaphor of our age, that of the computer, in an attempt to explain what he conjectures this might entail. He defines text processing as "skilled problem-solving" and attempts to explain how the individual links their knowledge and skills in decoding a text. Initially he proposes a mirroring of the receptive and active skills of reading and writing. The reader interacts with a text applying linguistic knowledge of syntax, semantics and pragmatics in order to ascertain the nature of the propositional content, illocutionary force(s), and text-type category presented. Essentially, the reader is seeking the answers to three questions:

- ✲ What is the text about?
- ✲ What was the writer's purpose?
- ✲ What plausible context could this test be used in?

Text-processing is characterized as being both bottom-up and top-down, as cascaded, and as interactive in as much as readers look for feedback loops on their understanding of the text, and are constantly revising their understanding of it. Movement, says Bell, is between a series of levels of text components (Table 5—2).

Table 5—2 Bell's levels of text components (1991)

Surface text
Linear sequences
Grammatical structures
Propositions
Sequencing devices
Essential ideas
Objectives, goals, and plans

The principles that Bell suggests are behind the mental processes involved are those of efficiency, effectiveness, and appropriateness, in that the reader must needs strike a balance in the degree of effort used in order to achieve their objective. Viewed from the reader/translator's perspective, Bell describes what is essentially the same model as Reiss. The differences between their proposals are in the emphasis Bell puts on the propositional content of the text, and on what he terms the "threshold of termination" for Text 1. This is the result of the author's production process, and is situated within a specific sociocultural, temporal and geographical context, The threshold for Text 2, is the text which results from the reader's interpretation process. Both of these are realized in the SL of the text, and the "threshold" represents the intentionality and illocutionary force of the author's Text 1, and the acceptability and perlocutionary force of the reader's Text 2. These parallel the relationship previously established between the author, implied author, implied reader and reader, established by Booth (1961) and Iser (1974). Again, Bell, like Reiss, Booth and Iser, sees the text as a communicative event, a product, fulfilling a role within a process. While his model draws on psycholinguistics and text linguistics, the basic relationships described are the same.

Bell looks at the individual's communicative ability as a system of resources which, like a database, can be drawn upon in order to participate in the communication process. Two key, and complementary elements in this process are defined by Beaugrande & Dressler (1981): intentionality, on the part of the author of the text; and acceptability, on the part of the reader. Intentionality is to do with the author's ability to produce a text that adopts the linguistic form most appropriate to achieve their communicative goal, in line with the psycholinguistic plan behind their communication. This is fully in line with the offer of communication and the act of communication described by Reiss in her model. Acceptability describes the inferences the reader is able to make about the text. It depends on the correlation the reader is able to perceive between the text as a discursive act and its application in the current situation. These inferences will be drawn from implicit or explicit references within the text, and the reader's acceptability of the text will be subject to a process of negotiation of propositional content as much as of linguistic content. This equates with Bell's definition of the decoding process as "skilled problem-solving", and further emphasizes the active, participative role of the reader in text processing, and the divisible, analytic nature of reading.

5.4 Qualitative resear ch into the translation process

Empirical research into the translation process based on the analysis of transcribed oral translations has, and is still being carried out by Wolfgang Lörscher (1986, 1992a, 1992b). The most significant fact about his work, we believe, is that it lends qualitative empirical weight to the theoretical arguments put forward by both Bell and Reiss, which we have described above.

Lörscher's initial position is that the phenomena of natural translation and of bilingualism give rise to a series of abilities which learned translation might enhance. He queries the nature of translation carried out by bilinguals, and some of the generalizations about bilingualism, which have led to the belief that it can be measured as an absolute condition, rather than being a point along a continuum.

He carried out this research with samples taken from three groups: advanced LB (English) learners, professional translators, and bilingual children. His objective is to elicit, describe, and compare the strategies employed in translating by subjects representing each of these three groups.

Lörscher's based his method on the use of transcriptions of oral translations from the subjects LA (German) into their LB (English). He derived his analysis from the first phase of the study, in which he developed a generative description of the processes that he interpreted as having occurred in the corpus that he had collected from his LB learners. This taxonomy of strategies he then used and refined in analyzing the oral translations of the second group, the professional translators. At this stage, he was able to develop a flow-charting approach that he used to describe the recurrent patterns in the translation process.

Significant results drawn from the initial phase of his study, and from the work he has carried out subsequently, reveal the following principal conclusions:

- Translating is a retrospective-prospective process
- The translation process is largely controlled by an expectation structure
- This expectation structure is built up mainly by
    - Separating SL forms from their meaning
    - Ideas about an "optimal" TL text
- Two extreme approaches to translation: sign-oriented and sense-oriented translation
- a marked similarity in the range of strategies employed by untrained translators and by the professional counterparts
- a divergence in terms of the frequency and distribution of the strategies.

These conclusions link in with the theoretical weight we have already seen attached to the reading process. The retrospective-prospective process is inherent to reading; expectation structures in reading are to a large extent responsible for interpretation of text, at the micro level of collocation, and at the macro level of text type or genre. Furthermore, the two processes he hypothesizes as being part of the expectation structure are both dependent on reading.

5.4.1 *DISCUSSION*

Some of the more questionable aspects of Lörscher's work are his basic research premise that the use of translation into the foreign language is essentially more revealing than that of translation into the mother tongue. Similarly, the quantitative aspects of his research lack a degree of depth. He is dealing with relatively small samples — the LB (English) data comes from 15 subjects (1986:277) — and consequently we can only generalize with caution.

In terms of the skills and sub-skills that Lörscher's subjects would appear to have been using in their translation processes these are highly analytical, and focus on the strategies he has perceived to be operating. Could these strategies be classified? We believe that a taxonomy of reading skills and sub-skills, such as that presented by Munby (1978), would help us to make such as classification. However, it would still lack an empirical base, remaining partially intuitive, as was Carroll and Sapon's approach to the MLAT (Chapter 4).

5.5 A process-oriente d approach to translator training

In this part of the chapter, we will study how two specialists in the pedagogy of translation view the underlying skills would-be translators need to acquire or develop as a part of their training. In particular, we will look at the links we may find between their affirmations and the conclusions of the research we have already analyzed.

5.5.1 *A SEQUENTIAL MODEL OF TRANSLATION*

Gile (1995) transfers the theoretical principles underlying translation as a process, described by Bell, Reiss and Lörscher, to the practical realities of translation pedagogy. He presents in top-down fashion a series of principles, methods and classroom procedures for translator training.

The basic concepts and models derive from Gile's experience as much as from empirical research. While he refers to a number of articles and books in support of specific positions, ideas, or concepts, he also affirms that he is drawing on observational experience. On a number of occasions, he clearly qualifies his proposals by highlighting the need for empirical data. His process-oriented approach is primarily a pedagogical instrument.

Within this pedagogical approach, Gile discusses the role of comprehension as a component of the translation process:

> *Other than statements stressing the very central role of comprehension in interpretation and translation, there have been few efforts to investigate scientifically the nature and extent of comprehension in I[nterpreting]/T[ranslation]. (75)*

Gile believes that within discourse comprehension the combination of world knowledge, which he calls extralinguistic knowledge, and analysis is especially important. The relationship between the two is complementary and there is a compensatory factor that ensures the one balances the other. We analyze words from the perspective of the co-text to resolve questions of polysemy, and world knowledge resolves questions of context.

Gile contrasts world knowledge with sociolinguistic knowledge, and with knowledge of cultural aspects of the two language communities. He suggests that the distinction between knowledge of the language and extralinguistic knowledge is not always easy to make, in particular as regards sociolinguistic and other cultural aspects of the langauge used in the relevant communities. Clearly, in order to use appropriate forms of politeness in a given situation, one needs to know not only linguistic rules, but also something about the particular production situation, which is extralinguistic. However, whether such elements are considered part of the knowledge of language or of extralinguistic knowledge, the basic relation represented by the comprehension equation remains valid (79).

In order to develop this further, Gile describes the comprehension equation:

$$\text{Comprehension} = \begin{array}{c}\text{Knowledge}\\ \text{of language}\end{array} + \begin{array}{c}\text{Extralinguistic}\\ \text{knowledge}\end{array} + \text{Analysis}$$

Among the models that Gile proposes is the Sequential Model of Translation. By design, this has a practical, didactic function based on the translation process and offering opportunities for error analysis. The model describes two completely distinct phases within the translation process: comprehension and reformulation. The author's suggestion is that by rigorously maintaining this division translators are able to involve themselves in more analysis and to reduce the amount of SL interference in the translation process.

Within this model, Gile describes criteria for use as iterative tests:

Plausibility: "... the translator looks at the idea or information he or she believes that the Translation Unit expresses, and examines it critically with respect to other information possessed so as to detect contradictions." (103)

Fidelity: "... he or she checks that none of the information has been omitted in the translation, and that no unwarranted information not contained in the source-language Translation Unit has been added..." (104)

(Editorial) Acceptability: "... he or she checks that in terms of clarity, language correctness, stylistic appropriateness, and terminological usage, it is acceptable." (104)

5.5.1.1 *Discussion*

The acknowledged lack of an empirical base to Gile's theories and methodological approaches to translation training clearly detract from the theoretical value we can give to his conclusions. However, the level of importance that he gives to the question of comprehension, to the analytical processes based on reading, and the way in which he describes the skills (lacking) in translators in training, lend further credence to our view that reading skills can serve as a means of measuring aptitude for translation.

5.5.2 THE TRANSLATOR "AS EXPERT READER"

Beeby Lonsdale (1996) makes one of the most recent contributions to the practice of translation teaching. She adopts the theoretical framework described by Bell (1991), and in her work, takes his three-part division of translation into theory, process, and text — which we have described earlier — and then develops each of these components. Her aim in doing so is to apply them to the practical questions of what she terms "prose translation" — which we call Translation A-B — from Spanish into English.

In our reading of Beeby Lonsdale's work we have encountered important insights into some of the unresolved problems arising from our study of the context. As throughout the history of aptitude testing, the question of whether or not to take an analytic or a synthetic approach to the testing and marking of a test of aptitude for translation depends very much, on how we see the nature of the translation process. Beeby Lonsdale makes a point of emphasizing, among other factors, the creative role of the reader, also in line with Booth (1961), Iser (1974), and Reiss (1992).

In discussing translation from the point of view of text, Beeby Lonsdale divides text into its macro- and microstructures at the pragmatic, semantic, syntactic, lexical, formal, and finally semiotic or intertextual levels. Throughout she is at pains to emphasize that the distinctive levels of structure described cannot be considered independently in the reality of the translation process. They are "cascaded", with constant movement taking place between micro- and macrostructure, in bottom-up and top-down directions. This attitude towards the levels of structure indicates that an analytic approach to the underlying skills employed in the translation process, and most specifically to the reading skills, may well be appropriate from a theoretical point of view. What is important, though, is the explicit recognition that the reality of the application of these skills will be much more complex than this implies.

Beeby Lonsdale describes the development of communicative translation theory by Delisle (1980, 1993) to deal with pragmatic texts; taking the translation unit to be the text as discourse; and from this communicative theory deriving a communicative teaching methodology. The theoretical framework, then, leads to the application of a variational approach based on the experience of translation in practice and as a process.

In the context of describing the teaching/learning process, she uses Bell's expression "the cycle of inquiry" (1991) in order to emphasize the need for translation training to be both inductive and deductive at the same time.

Beeby Lonsdale situates the active role of the translator as reader in that created by Barthes (1975) and developed more specifically by Bassnett-McGuire (1980:79) and Nord (1991:16).

She makes the point that

> *Pre-translating text/discourse analysis provides not only a full comprehension of the SLT/discourse and an explanation of its linguistic, textual, and discourse structures and their relationship with the system and rules of the SL and the SL culture, but also a reliable basis for every decision that the translator has to make in a particular translation process. Therefore, text/discourse analysis for the translator has to be integrated into a general theory of translation that will serve as a frame of reference. (50)*

Pre-translation exercises of reading comprehension are

> *to develop their [the learners'] knowledge of context through reading and/or to make active use of their own knowledge. (50)*

Learners make errors in meaning due to insufficient knowledge of the LA and/or failure to activate their existing knowledge. These errors lead to elements of intertextuality being lost, or ignored because learners have simply not read sufficient parallel texts to be able to recognize them.

Beeby Lonsdale subdivides the knowledge that trainee translators possess into four categories:

Table 5—3

Situational	which comprises knowledge of places, people and events
Verbal	including an awareness of idiom and the ability to use idiom as the key to resolving issues of polysemy
Socio-historical	an awareness of events, codes and relationships from a real-world perspective
Cognitive	to be able to "collate" the information presented in the text

Based on the previously described outline of communicative competence, which is derived from Canale (1987), and Bell (1991), Beeby Lonsdale offers her own version of Ideal Translator Communicative Competence (Table 5—4). Essentially this follows Canale, doing away with Chomsky's distinction between competence and performance and including both knowledge and skills in the use of that knowledge under the same heading. She fails to go into any detail on the way in which these elements are related but, rather, lists the four components and defines each in turn.

Ideal translator communicative competence, she then contrasts, in anecdotal terms, with the reality of those students she teaches. The profile she offers is brief and sketchy, but it does appear to match our own student profile in Granada (Table 5—5). Taken one by one, Beeby Lonsdale offers a description of each of the four competences with regard to SLTs and TLTs.

Table 5—4 Ideal translator communicative competence (Beeby Lonsdale 1996:92)

Ideal translator grammatical competence.	Knowledge of the rules of both languages, including vocabulary and word formation, pronunciation, spelling and sentence structure-that is, the knowledge and skills required to understand and express the literal meaning of utterances.
Ideal translator sociolinguistic competence.	Knowledge of and ability to produce and understand utterances appropriately in the situational context of both cultures-that is, as constrained by the cognitive context, the general sociohistorical context, the mode, the field, the tenor, the status of the participants, the purposes of the interaction, the skopos of the translation, and so on.

Ideal translator discourse competence.	The ability to combine form and meaning to achieve unified spoken or written texts in different genres in both languages. This unity depends on cohesion in form (the way in which utterances are linked structurally to facilitate interpretations of the text) and coherence in meaning (the relationships among the different meanings in a text: literal meanings, communicative functions or social meaning, intertextuality).
Ideal translator transfer competence.	The mastery of communication strategies that allow transfer of meaning from the SL to the TL and may be used to improve communication or compensate for breakdowns (caused by limiting factors in actual communication or insufficient competence in one or more of the other components of communicative competence).

The translation process is made up of three stages: comprehension, deverbalization, and reformulation and verification.

Table 5—5 Beeby Lonsdale's student profile

Expert reading skills in the SL (grammatical, sociolinguistic, and discourse competence).	Achieving advanced reading skills in the SL and recognition of limitations in this area.

5.5.2.1 *Discussion*

In the course of her work, Beeby Lonsdale makes a number of assertions about the nature of translator communicative competence and, in Table 5—6, we link these to Munby's taxonomy of reading skills (1978), which we will describe in more detail in Chapter 6.

Table 5—6 Ideal translator communicative competence (Beeby Lonsdale 1996) contrasted with a taxonomy of communicative reading skills (Munby 1978)

Translator communicative competence	*Reading skills*
Different languages organize meaning and lexis in different ways. Semantic fields are rarely exactly equivalent-for example, *correr* and run.	19, 32
Lexical polysemy is resolved by context. E.g. double the money, double the blanket, he has a double...	19, 22
Syntactic polysemy is resolved by context. E.g. his car, his house, his arm...	28, 32
Collocation is not rule-based.	30.6
Standardized language must be distinguished from non-standardized language.	21, 26
Context affects register (field, mode, and tenor).	19.2, 22.1
Negotiating meaning requires awareness of pragmatic purpose and intertextuality.	34, 35, 37.4
Discourse cohesion and coherence are expressed differently by different languages.	47-54

We believe that these elements of ideal translator communicative competence — elements that need to be acquired in the training process — coincide with the reading skills and sub-skills that Munby has included in his taxonomy.

We propose to adopt these skills and sub-skills as a set of operational definitions that will serve as the target skills we will test in our entry test in aptitude for translation. In Chapter 6, we present Munby's work more fully, and detail the skills we have selected as target skills for inclusion in our first draft test specification.

5.6 Conclusions

From our research into translation theory and translation pedagogy we have reached the conclusion that reading skills play a significant part in translation, when this is considered as a complex, analytical process. This is particularly important from the pedagogical point of view.

We believe that translation is a process susceptible to analysis, and that aptitude for translation is a compound of many skills, and that these can be described and analyzed on individually. In Chapter 6, we will look at Munby's taxonomy of reading skills and sub-skills (1978), which provides definitions of these skills that are equivalent to the elements of translator communicative competence described as essential by specialists in translation pedagogy.

6 THE STARTING POI NT FOR OUR TEST DESIGN: MUNBY, AND THE DESCRIPTION OF READING SKILLS

6 THE STARTING POINT FOR OUR TEST DESIGN: MUNBY, AND THE DESCRIPTION OF READING SKILLS

6.1 THE DESCRIPTION OF READING SKILLS

6.1.1 *Why test in LA and LB?*

6.2 CONCLUSIONS

IT IS NOWADAYS a commonplace to lament the lack of a recent alternative to the work carried out by Munby in preparing his *Communicative Syllabus Design*, published as long ago as 1978 (Spolsky 1995). This in no way is a criticism of his taxonomy of communicative language skills, but rather a tribute to an author whose work has yet to be seriously challenged. Consequently, and in the light of our analysis of pertinent research, Munby's work is essential to this study.

In this chapter, we will describe some of the background to Munby's work and present the sub-set of skills that we have selected for our test specification.

Munby began from the definition of English for Specific Purposes (ESP) that was gaining currency in the mid- and late seventies. This distinguished ESP from General English by virtue of the learner-centered nature of the syllabus design. An ESP, or indeed any other course designed for "specific" purposes – English for Academic Purposes (EAP) for example – was intended to be much more closely tailored to the needs of the learners than were other "general" courses. The essence of this tailoring was the application of a needs analysis approach to the definition of the learners' communicative needs. The products of this would then become the basis of syllabus content. Munby (1978) was, and today remains, the first and foremost attempt to produce an instrument capable of deriving that list of learner-centered communicative needs.

The origins of Munby's model lies in work carried out in the early 70s. On the one hand, Hymes (1971) and Halliday (1974) within the field of sociolinguistics had described aspects of sociolinguistics competence which began to offer linguists a tool with which to measure individual language performance. Simultaneously, in Britain and Europe, linguists working within Foreign language learning (FLL), were trying to define and describe micro-skills which could form the basis for a widely accepted specification of language use in order to achieve comparability in the language learning programmes used throughout Europe. The product of this research was the "Draft outline of a European unit/credit system" (Trim 1973) which was the precursor of the Threshold level (van Ek 1975), and the subsequent developments of other levels, both more advanced and more elementary. Trim, and his colleagues at the Council of Europe, spawned a number of milestone publications within the FLL world (e.g. Wilkins 1976), many of which contributed to a veritable revolution which found its way into all aspects of FL learning and teaching.

The aim of Munby's work was to design a "dynamic processing model" which would offer linguists with a systematic approach to the analysis of the learner's needs such that communication needs might be established. From this list, Munby proposed, the course designer might able to draw up the "target communicative competence" required by the learner — as an individual, or as a "type" — and thus produce a solid foundation on which to build a syllabus. Munby expressly rejected any attempt to involve his model in the latter part of this process.

The list presented below is the selection of skills on which our entry test in aptitude for translation studiest was based. It forms the starting point of the test design process.

Table 6—1 Reading skills and sub-skills selected for our test in aptitude for translation. Numbering corresponds to Munby's original classification. (Munby 1978:126-131)

- *Deducing the meaning and use of unfamiliar lexical items, through (Skill 19)*
  - understanding word formation (Skill 19.1)
    - stems/roots (Skill 19.1.1)
    - Affixation (Skill 19.1.2)
    - Derivation (Skill 19.1.3)
    - Compounding (Skill 19.1.4)
  - contextual clues (Skill 19.2)
- *Understanding explicitly stated information (Skill 20)*
- *Understanding information in the text, not explicitly stated, through (Skill 22)*
  - making inferences (Skill 22.1)
  - understanding figurative language (Skill 22.2)
- *Understanding conceptual meaning, especially (Skill 24)*
  - quantity and amount(Skill 24.1)
  - definiteness and indefiniteness(Skill 24.2)
  - comparison; degree(Skill 24.3)
  - time (esp. tense and aspect)(Skill 24.4)
  - location; direction(Skill 24.5)
  - means; instrument(Skill 24.6)
  - cause; result; purpose; reason; condition; contrast(Skill 24.7)
- *Understanding the communicative value (function) of sentences and utterances (Skill 26)*
  - with explicit indicators (Skill 26.1)

without explicit indicators (Skill 26.2)

e.g. an interrogative that is a polite command; a statement that is in fact a suggestion, warning, etc. depending on the context; relationships of result, reformulation, etc., without "therefore", "in other words", etc.

Understanding relations within the sentence, especially (Skill 28)

elements of sentence structure (Skill 28.1)
modification structure (Skill 28.2)
Premodification (Skill 28.2.1)
Postmodification (Skill 28.2.2)
Disjuncts (Skill 28.2.3)
negation (Skill 28.3)
modal auxiliaries (Skill 28.4)
intra-sentential connectors(Skill 28.5)
complex embedding (Skill 28.6)
focus and theme: (Skill 28.7)
Thematic fronting; and inversion (Skill 28.7.1)
Postponement (Skill 28.7.2)

Understanding relations between parts of text through lexical cohesion devices of (Skill 30)

repetition (Skill 30.1)
synonymy (Skill 30.2)
hyponymy (Skill 30.3)
antithesis (Skill 30.4)
apposition (Skill 30.5)
lexical set/collocation (Skill 30.6)
pro-forms/general words (Skill 30.7)

Understanding relations between parts of text through grammatical cohesion devices of (Skill 32)

reference (anaphoric and cataphoric) (Skill 32.1)
comparison (Skill 32.2)
substitution (Skill 32.3)
ellipsis (Skill 32.4)
time and place relaters (Skill 32.5)
logical connectors (Skill 32.6)

Interpreting text by going outside it, (Skill 34)

using exophoric reference (Skill 34.1)
"reading between the lines" (Skill 34.2)

integrating data in the text with own experience or knowledge of the world (Skill 34.3)

Recognising indicators in discourse for (Skill 35)

introducing an idea (Skill 35.1)

developing an idea (e.g. adding points, reinforcing argument) (Skill 35.2)

transition to another idea (Skill 35.3)

concluding an idea (Skill 35.4)

emphasising a point (Skill 35.5)

explanation or clarification of point already made (Skill 35.6)

anticipating an objection or contrary view (Skill 35.7)

Identifying the main point or important information in a piece of discourse, through (Skill 37)

vocal underlining (e.g. decreased speed, increased volume) (Skill 37.1)

end-focus and end-weight (Skill 37.2)

verbal cues (e.g. 'The point I want to make is ...')(Skill 37.3)

topic sentence, in paragraphs of (Skill 37.4)

Inductive organisation (Skill 37.4.1)

Deductive organisation (Skill 37.4.2)

Distinguishing the main idea from supporting details, by differentiating (Skill 39)

primary from secondary significance (Skill 39.1)

the whole from its parts (Skill 39.2)

a process from its stages (Skill 39.3)

category from exponent (Skill 39.4)

statement from example (Skill 39.5)

fact from opinion (Skill 39.6)

a proposition from its argument (Skill 39.7)

Extracting salient points to summarise (Skill 40)

the whole text (Skill 40.1)

a specific idea/topic in the text(Skill 40.2)

the underlying idea or point of the text(Skill 40.3)

Selective extraction of relevant points from a text, involving (Skill 41)

the coordination of related information (Skill 41.1)

the ordered rearrangement of contrasting items (Skill 41.2) the tabulation of information for comparison and contrast (Skill 41.3)

Expanding salient/relevant points into summary of (Skill 42)

the whole text (Skill 42.1)

a specific idea/topic in the text (Skill 42.2)

Reducing the text through rejecting redundant or irrelevant information and items, especially (Skill 43)

Omission of closed-system items (e.g. determiners) (Skill 43.1)

Omission of repetition, circumlocution, digression, false starts (Skill 43.2)

Compression of sentences or word groups (Skill 43.3)

Compression of examples (Skill 43.4)

use of abbreviations (Skill 43.5)

use of symbols denoting relationships between states, processes, etc. (Skill 43.6)

Basic reference skills: understanding and use of (Skill 44)

Graphic presentation, viz. headings, sub-headings, numbering, indentation, bold print, footnotes (Skill 44.1)

table of contents and index (Skill 44.2)

cross-referencing (Skill 44.3)

card catalogue (Skill 44.4)

Phonemic transcription/diacritics (Skill 44.5)

Skimming to obtain (Skill 45)

the gist of a text (Skill 45.1)

a general impression of the text (Skill 45.2)

Scanning to locate specifically required information on (Skill 46)

a single point, involving a simple search (Skill 46.1)

a single point, involving a complex search (Skill 46.2)

more than one point, involving a simple search (Skill 46.3)

more than one point, involving a complex search (Skill 46.4)

a whole topic (Skill 46.5)

Transcoding information presented in diagrammatic display, involving (Skill 51)

- Straight conversion of diagram/table/graph into speech/writing (Skill 51.1)
- Interpretation or comparison of diagrams/tables/graphs in speech/writing (Skill 51.2)

Transcoding information in speech/writing to diagrammatic display, through (Skill 52)

- Completing a diagram/table/graph (Skill 52.1)
- Constructing one or more diagrams/tables/graphs (Skill 52.2)

Recoding information (expressing/understanding equivalence of meaning)(Skill 53)

- within the same style (e.g. paraphrasing to avoid repetition) (Skill 53.1)
- across different styles (e.g. from technical to lay) (Skill 53.2)

Relaying information (Skill 54)

- Directly (commentary/description concurrent with action) (Skill 54.1)
- Indirectly (reporting) (Skill 54.2)

The virtues of Munby's taxonomy can be seen in the number of subsequent authors who have employed them in their own work. Grellet (1981), Madsen (1983) and Heaton (1988), all use variations of the basic terms defined by Munby, although they tend to simplify the terms, and to gloss skills and subskills in more teacher-friendly language (e.g. Grellet 1981:4-5).

In an unpublished PhD thesis on testing for EAP students, Weir (1983) modified Munby's taxonomy of reading skills by introducing a distinction between lower and higher order skills:

Table 6—2 Lower order reading skills (Weir 1983)

Reference skills, eg using bibliography, index, footnotes
Deducing the meaning and use of unfamiliar lexical items through understanding word formation and contextual clues
Understanding relations within the sentence
Understanding relations between parts of text through grammatical cohesion devices
Understanding relations between parts of text by recognising indicators in discourse

Understanding the communicative function of sentences, with and without indicators
Understanding conceptual meaning, eg cause, result, purpose
Understanding explicitly stated ideas

Table 6—3 Higher order reading skills (Weir 1983)

Understanding ideas not explicitly stated
Separating essential from non-essential in text, distinguishing the main idea from supporting details, etc
Transfer of information from one medium to another
Skimming
scanning for specifics
surveying to obtain gist
Notemaking
selective extraction of relevant and related points for summary
extracting salient points for commentary
reducing text by rejecting redundant or irrelevant items
Critical evaluation

As can be seen, the terminology Weir uses is essentially the same as Munby had used originally, but, as have others he has simplified slightly the descriptions.

6.1 The description of reading skills

The decision to focus one part of our entry test of aptitude for translation on reading skills is based on two premises. The first is the theoretical basis that we have described in the Chapter 4, and the second is pragmatic, and also based on the information derived from our study of the context (Chapter 3) in which we were to work.

As we have discussed earlier, it was felt that the focus of such a test had to be based on an understanding of the learning requirements of students actually studying to become translators. Having studied work on the process structures believed to underpin translation, we found that the conclusions of both Translation studies theorists and specialists in the pedagogy of translation, coincided with the wealth of data we had available from the diploma course students, and their largely undocumented history of failure. It was clear that here was a link between the two. The question that had to be asked was "Which skill(s) were those who passed the course able to employ that the remainder were not?" This approach led to a further analysis of failure and of the teaching that brought about recovery among these learners.

The course final examination that these candidates were unable to perform well in — which we have discussed in great detail earlier — involved a summary writing exercise. The results showed that their inability to produce coherent summaries of the texts offered correlated highly with their overall failure in the examination. The conclusion we drew was that

- the inability to read in depth, expressed as a lack of the ability to consciously apply LA or LB reading skills indicated a crucial weakness in this area.

Students writing skills appeared not to be in question as their performance in essay writing was markedly better.

As we have mentioned before, in 1991-92, the course programme for *Traducción I (Inglés)* was rewritten to focus on the development of specific reading skills, and on the practice of writing skills in conjunction with these. Texts were analysed in such a way as to encourage learners to develop their ability to consciously develop a systematic approach to reading. The mirroring of this analytical approach was then applied to their development of writing skills.

In addition, because of the constraints facing the administration of the test in terms of the numbers of candidates it was believed that an objective test which could eventually be corrected by computer was far preferable to any subjectively marked test of the kind which a focus on writing skills would entail. We will take up this aspect of the study further in the next chapter, when we look at the issues derived from testing theory, which impinge on our study.

6.1.1 *WHY TEST IN LA AND LB?*

The attitude of colleagues in the University of Granada towards the question of an entry test to the degree program was very clear: the most important stumbling block to producing quality graduates would be the level of Language B with which learners entered the program. This attitude is contrary to the research we have reported on in Chapter 3 and to that expressed throughout the francophone world in translation studies (Bossé-Andrieu 1981). It seems to this writer to stem very much from a suspicion of the quality of foreign language teaching in the Spanish state system. Again, this position was unsupported in research in the literature, although the failure rate we have described earlier clearly indicated something.

The teaching of the first year translation course had been in the hands of professional translators, with some, little, or no experience or training in the teaching of foreign languages. The standards and criteria that they employed were, then, derived from their professional background, and these led to the high failure rate and their conclusion that this was due to 'poor knowledge of the foreign language'.

The issue of the relationship between reading ability and language proficiency, whether in LA or in LB is "equivocal and tentative" (Alderson 1984:24). In broad terms, it can be said that some sort of relationship has been demonstrated to exist between LA reading ability and LB reading ability, and that this relationship enjoys a degree of independence from LB proficiency according to the level of that proficiency. Alderson, in his survey of the subject, finds support for the belief that some type of "threshold" level of LB proficiency needs to be attained by individuals in order for their LA reading abilities to influence their LB reading.

However, he makes the complexities of the relationship, and the need for further qualitative research obvious, and leaves us in little doubt that, in our study, the importance of reading skills in both LA and LB is such that we cannot afford to test one only.

6.2 Conclusions

Munby's taxonomy of communicative language skills has provided us with a basic list of individual reading skills that we believe are equally applicable to all of the languages involved in our entry test procedures (Spanish, English, French and German). Following on from our decision to adopt the concept of aptitude as a divisible entity, these skills are discretely defined in such a manner that they offer clear targets for us to aim at when we write individual test items. Furthermore, these individual test items can be grouped together in such a way as to offer a range of target skills in our sub-tests. The only difficulty that we foresee lies in the complex nature of our target, as described in Chapter 5, which may mean that the overlaps between different aspects of aptitude confuse aspects of our results.

7 Classical Test Th eory

7 CLASSICAL TEST THEORY

7.1 Major trends in testing
7.2 Classical Test Theory
7.2.1 *Validity*
7.2.1.1 Backwash validity
7.2.1.2 Concurrent validity
7.2.1.3 Construct validity
7.2.1.4 Content validity
7.2.1.5 Criterion-related validity
7.2.1.6 Face validity
7.2.1.7 Predictive validity
7.2.1.8 Response validity
7.2.1.9 How valid is valid?
7.2.1.10 Conclusions
7.2.2 *Reliability*
7.2.2.1 Internal consistency
Split-half estimates
Item variance reliability estimates
7.2.2.2 Stability
7.2.2.3 Spearman-Brown Prophecy Formula
7.2.2.4 Standard error of measurement (S_e)
7.2.3 *Discussion: Reliability vs. Validity*
7.2.4 *Practicality*
7.3 Facility value
7.4 Discrimination
7.5 An alternative to CTT?: Item response theory
7.6 Conclusions

THE OBJECTIVES of this chapter are

- to provide an overview of the principle issues in Classical Test Theory (CTT) as it is applied to Foreign Language Learning
- to define the basic theoretical approach to testing that we have chosen for our research
- to discuss the statistical formulae we consider appropriate to achieve our objectives.
- to define those concepts drawn from CTT that we believe have an important bearing on our research

To do this, we present an overview of the main considerations drawn from the field of testing that have a direct or indirect influence on our study. We begin with a look at the major trends in testing in recent years, and outline the fundamental arguments that have fuelled debate among practitioners. Then we define validity, reliability, and practicality, three crucial concepts discussed throughout the literature, and highlight aspects of each that affect our research. We next distinguish between Classical test theory (CTT) and Item response theory (IRT), these being the two fundamental approaches to the statistical analysis of test scores. In this study, we will only apply techniques drawn from the former. However, current research in IRT, dubbed "Modern test theory" by some (e.g. Crocker and Algina 1986), makes it clear that these two theories are to be considered complimentary approaches to the subject. Further research into testing for aptitude would need to adopt the statistically more complex IRT method in order to broaden even further the range of information available.

In the final sections of the chapter, we define the most important CTT statistics that we will calculate from our test results, and indicate the ways in which these can be interpreted. The measures in question are the FACILITY VALUE (FV), and DISCRIMINATION INDEX (DI). Both of these provide information on the performance of individual items; as do the descriptive statistics, and the frequency distribution of scores, which provide information on the performance of the test as a whole.

7.1 Major trends in testing

We are now going to give a brief sketch of some of the major trends in testing practice that have been described in the literature as a backcloth against which to view elements of our research design and methodology.

The three principal stages in the development of language testing are generally labeled:

❶ Psychometric-structuralist,

❷ Psycholinguistic-sociolinguistic,

❸ communicative

(Spolsky (1976) quoted by Weir (1988:1).

They are considered to represent a chronological sequence of developments, although there has been considerable overlap between them. In these three periods, researchers have been involved in a tense debate over the primacy of test reliability or that of test validity. During the so-called psychometric-structuralist era, reliability held sway with the application of DISCRETE ITEM tests, many of which were based on the multiple-choice model, and which were highly objective.

As test theory developed, the INTEGRATIVE TEST gained ground with the use of dictation and the cloze test, both of which moved away from the isolated testing of language components to examining aspects of linguistic competence. However they did not require the use of spontaneous language production, and they did not test communicative competence. Moreover, the debate continued over the nature of linguistic competence, whether this was divisible and therefore susceptible to being tested in discrete item tests, or whether it was unitary and, accordingly, best tested through integrative test tasks.

The major criticism of the integrative tests was that they too, failed to test the spontaneous use of language, and therefore did not demonstrate the individual's ability to communicate. With the development of Communicative language learning and teaching, the drive began to discover a communicative approach to language testing.

Against the background of advances in language testing theory, in the development of tools and instruments of statistical analysis, and in the application of the principles of Communicative language learning and teaching, Bachman (1991) describes a theoretical framework based on a multicomponential model of language ability. He applies this to the design of test tasks to demonstrate its applicability to the Communicative language testing philosophy.

Language ability, Bachman suggests, comprises two components: language knowledge and metacognitive strategies. The first, also termed competence, is exclusive to language use, and divides into organizational knowledge and pragmatic knowledge. Of these, the former includes how texts are organized and is made up of grammatical knowledge and textual knowledge; the latter deals with the elements of sentences, intentions and contexts, and the ways in which they relate to form meaning: propositional, functional and sociolinguistic knowledge. Alongside these, Bachman presents a set of metacognitive strategies termed Assessment, Goal-setting, and Planning. He suggests these interact simultaneously with each other, and with the components of language knowledge. Test task performance, he states, reflects all of these factors, and the interactions between them.

In Bachman's opinion, the question of test-taker characteristics, and the interaction between the test-taker and the test task comes to the fore as the "interface" between SLA and language testing research. This interaction influences test scores, and is clearly relevant to test activity design.

Now it is Bachman's contention that validity remains the more important of the two criteria by which we judge tests, but as reliability is understood to be one essential part of validity, this is never ignored.

Bachman's contribution to the current phase of language testing development places the learner or in this case the test-taker, at the center of the process.

Table 7—1 Bachman's parameters of authenticity (1990)

		Interactional authenticity	
		Low	High
	High	1	4
Situational authenticity			
	Low	3	2

In Table 7—1, we can see the frame that Bachman proposes we use in order to judge the authenticity of the test task. If we consider that a communicative test task should have real-world authenticity, then we must judge this against two scales: interactional authenticity, and situational authenticity. These scales represent high-low ranges, and the positions numbered 1 to 4 identify extremes. Our sub-tests in reading skills for translators would, we think, correspond to position 1. That is, they are low on interactional, but high on situational authenticity. We believe this to be so because the reading phase of the translation process involves no direct communicative interaction, certainly no in the terms in which the FL practitioner would recognize. However, it does have high situational authenticity, in that it is an essential part of the reality of the translation process. After all, the translator is the "expert reader" (Beeby Lonsdale 1996). The questions arising from the interpretation of the reading process are dealt with in more detail elsewhere (insert reference).

We feel that, while the sub-tests we apply lack the some of the essential characteristics that from an FL point of view would characterize them as communicative, they are in fact more communicative, in the narrow context of testing for aptitude for translation than they at first appear. The initial standpoint from which we view the process of reading, namely that of an interactive, communicative process, strengthens the case for judging our sub-tests against Bachman's criteria for authenticity.

1.2 Classical Test Theory

In this section, we will look at the essential equation from which Classical test theory (CTT) derives its analytical procedures.

Table 7— 2 True score and measurement error

$$X = T + E$$

Classical test theory (CTT), also known as True Score Theory, deals with the relationship between test scores observed as the result of administering test instruments, and the "real" scores, that indicate individual ability. The basic model of CTT is described in Table 7—2.

Here, X represents the observed score; T, the true score the candidate would achieve taking a perfect test under perfect conditions; and E the part of the observed score that is due to error of one sort or another.

This theory presents the test user with a dilemma: while the observed scores can be manipulated for statistical analysis, how can we calculate the true score and the score due to error? The approach to resolving this is based on calculating test reliability and error of measurement, both of which we will discuss in more detail later.

Measurement error in the use of a test is responsible for some of the variation in test scores. The sources of error lie in different aspects of the test, and each of them is susceptible to statistical method, giving rise to the calculation of coefficients of reliability.

method, giving rise to the calculation of coefficients of reliability.

We will now define and describe the essential concept of validity, and draw a list of those aspects that we need to ascertain in our research process.

7.2.1 VALIDITY

The extent to which scores on a test enable inferences to be made which are appropriate, meaningful, and useful, given the purpose of the test. (ALTE 1998:168)

Whenever test users interpret scores achieved in a particular test as being indicative of "something", they need to know to what extent they can justify their interpretation (Bachman & Palmer 1996:21).

If, as is our case, the purpose of the test is to enable decisions to be made, then the validity of the test scores is paramount. Even more so when we are dealing with scores that will influence decisions over individuals' future studies and, indeed, professional careers. The scope of the inferences we may make from a test score is wide-ranging, and some of the more recent literature on the subject (Bachman 1990; Alderson, Clapham and Wall 1995) explicitly includes consideration of these important ethical and social dimensions of test use.

In general, validity is seen as an umbrella term, which is divided up in a number of ways. Writers tend to position themselves at different points along a continuum that can best be described in terms of the orientation and interests of their target readers. At one end of the scale we find authors such as Baxter (1997), Heaton (1988), Hughes (1989), and Weir (1990) who include classroom teachers among their target audience, and who aim to inform and influence classroom practice. Consequently they are more didactic and less statistical in their presentations. In their acknowledgment of the less didactic end of the scale, they refer readers to others, (e.g. Bachman 1990; Crocker and Algina 1986; Henning 1987). These are writers who take a more academic, research-oriented approach, and who aim to communicate detailed methods of analysis, offering much in the way of statistical formulae and procedures. All of them, however, take essentially the same position as to the nature of validity, classifying it in different ways as Table Testing/3 shows.

There is much basic agreement among these authors. Moreover, the disagreements tend to indicate their relative positions on the

Table 7 — 3 Validity and validities.

	Baxter	*Crocker and Algina*	*Heaton*	*Henning*	*Hughes*	*Weir*
Backwash	✓	✗	✗	✗	✗	✓
Concurrent	✗	✓	✓	✓	✓	✓
Construct	✓	✓	✓	✓	✓	✓
Content	✓	✓	✓	✓	✓	✓
Criterion-related	✗	✓	✗	✓	✓	✓
Empirical	✗	✗	✓	✗	✗	✗
Face	✓	✓	✓	✓	✓	✓
Non-Empirical	✗	✗	✗	✓	✗	✗
Response	✗	✗	✗	✓	✗	✗
Predictive	✗	✓	✓	✓	✓	✓

continuum. We will now look at each type of validity in more detail.

7.2.1.1 *Backwash validity*

> *...the effect that a final test has on the teaching program that leads to it.* (Baxter 1997:28)

The difficulties of measuring the effect of testing on classroom practice is one that has been described in quantitative terms by Alderson and Wall (1993). This aspect of testing is clearly of importance as tests within educational systems are modified because merely by effecting changes in the testing procedures, changes in teaching will be brought about. The exact nature of this relationship is much less clear than is often supposed. However, this does not affect our current research, so we will take it no further.

7.2.1.2 *Concurrent validity*

> *A test is said to have concurrent validity if the scores it gives correlate highly with a recognized external criterion which measures the same area of knowledge or ability.* (A.L.T.E. 1998:139)

As an empirical measure of the relationship between scores this statistic is of importance in a study such as ours. When we described "The Diploma experience", we gave details of the scores obtained when we administered the Oxford Placement Test (Allan 1990). We mentioned its stated reliability, and described its component sub-tests. In our study we will present the results of applying this test, the OPT, as a measure of establishing the concurrent validity of our LB sub-tests. We do not predict a high correlation between the OPT scores and our sub-test scores because they do not measure exactly the same areas of knowledge. However, the comparison will serve to establish whether or not our test is a minimally valid measure of English language skills.

7.2.1.3 *Construct validity*

> *A test is said to have construct validity if scores can be shown to reflect a theory about the nature of a construct or its relation to other constructs.* (A.L.T.E. 1998:139)

This concept relates the test and the underlying applied linguistic theory of the acquisition of abilities and skills on which it is based. The construct we test, and the test mechanism we use, must be in harmony to achieve construct validity. This implies that the Communicative approach to language learning and teaching, is incompatible with the use of multiple choice questions (MCQs) as the test mechanism because they lack a communicative dimension. In our case, where the focus of the test is on reading skills, MCQs could well be considered the most valid means of testing the construct, because they only involve reading.

7.2.1.4 *Content validity*

> *A test is said to have content validity if the items or tasks of which it is made up constitute a representative sample of items or tasks for the area of knowledge or ability to be tested.* (A.L.T.E. 1998:140)

Within a teaching program, content validity describes the relationship between the language being tested and the course objectives. In order to ensure this is the case test writers are recommended to produce a test specification giving detailed information on the skills and areas to be tested, and on the weighting applied to each (Carroll B.J. 1980; Carroll B.J. and Hall 1985; Heaton 1988). In this context, "weighting" means quantifying and balancing test components so that the importance given to each, reflects the importance they have been given during the teaching/learning period (Heaton 1988:161).

Content validity, then, is an intrinsic requirement of any test. However, as we are working with no direct contact to any teaching program we have to ask ourselves if our test includes a representative sample of the skills it sets out to measure, with reference only to our written test specification. The preparation of that specification is, therefore, crucial to the content validity of the test. If we are able to draw on teaching programs of the type that candidates are likely to have experienced, and of the type that successful candidates will follow, then we are likely to produce a specification which represents a close approximation to the representative sample that is called for.

The evaluation of content validity is often left to the judgement of Subject Matter Experts, who are asked to compare test items and specifications. This process can, in itself, lead to a range of responses. Alderson (1990) found that so-called "judges" agreed little as to the target skills that test writers believed they had focussed on. In contrast, Bachman (1990) reports that a satisfactory level of agreement can be achieved. In the case of our tests, the opinions of members of the Examining Board were sought, as were those of colleagues involved in the test-writing process.

7.2.1.5 *Criterion-related validity*

> *A test is said to have criterion-related validity if a relationship can be demonstrated between test scores and some external criterion which is believed to be a measure of the same ability.* (A.L.T.E. 1998:140)

Also called empirical or statistical validity by some writers (Heaton 1988; Henning 1987), this aspect of validity is the general term that covers the relationships between any one specific test and other external tests, which have previously been validated. These measures can be then be used as criterion measures against which to judge a new test. The means by which criterion-related validity is measured is the use of correlation coefficient formulae such as the PEARSON PRODUCT-MOMENT COEFFICIENT. The values of coefficients derived from these calculations are interpreted against labels, such as those we mentioned earlier in our description of "Bossé-Andrieu's study of francophone countries". A more precise interpretation is made with reference to tables of statistical significance, for example those found in Henning (1987) Siegel and Castellan (1988), and Silver (1997).

7.2.1.6 *Face validity*

> *The extent to which a test appears to candidates, or those choosing it on behalf of candidates, to be an acceptable measure of the ability they wish to measure.* (A.L.T.E. 1998:144)

Henning (1988) classifies face validity as a non-empirical form of validity, and as such gives it little relevance. It is usually tested qualitatively through follow-up questionnaires. We consider it an essential aspect of a test because it conditions the mental approach of candidates to that test. If a test lacks face validity, candidates may not take it seriously, or they may be disconcerted by it to the extent that they under-perform. In the current context, face validity of aptitude testing at the University of Granada is highly significant, as candidates are required to pay a substantial examination fee.

7.2.1.7 *Predictive validity*

> *An indication of how well a test predicts future performance in the relevant skill.* (A.L.T.E. 1998:157)

Predictive validity is the essential purpose of all aptitude tests. The fundamental question that our test is designed to answer is one of future performance in translation. Clearly, the most appropriate aptitude test would be one of translation, as that would comply with the requirement that we measure the "relevant skill". However, as we have mentioned in our introduction, the legal framework that established the aptitude test expressly prohibits the use of translation as a test instrument. Therefore, as in the case of concurrent validity, we would not expect the correlation coefficients between our test of aptitude and a subsequent measure of translation ability to be significantly strong.

7.2.1.8 *Response validity*

> *Response validity is the extent to which examinee responses to a test or questionnaire can be said to reflect the intended purpose in measurement. Lack of adequate instructions, incentives, task familiarity, or courtesy could invalidate responses.* (Henning 1987:196)

If a candidate is in some way confused, or led into error due to an unexpected test task type, or inadequate instructions, they will not necessarily respond in the way expected. This will invalidate the candidate's response, although they may well indicate they are able to answer the item correctly.

7.2.1.9 *How valid is valid?*

Despite relative agreement over the classification of validity, there is still much that is open to debate. Heaton makes it clear that overall validity – i.e. the sum of the types we have just described - and the testing situation are inextricably linked. Test purpose is always a key factor in the interpretation of results measuring these concepts, but more than this, we must remember that "...the established criteria for measuring validity are themselves very suspect..." (Heaton 1988:162). Many writers have researched this area, as we will now see.

Fulcher (1987) criticizes trends in communicative testing and in particular the work of the English Language Testing Service (ELTS) (Carroll B.J. 1980, 1981, 1985; Munby 1978). He specifically finds fault with the search for content validity at the expense of construct validity in oral tests, and concludes that the piecemeal assessment of skills and sub-skills does not match the reality of authentic discourse. Matthews (1990) also questions oral proficiency testing, based on the use of criteria rating scales from the standpoint of an assessor. She believes that the search for content validity has led to an over-reliance on discrete sub-skills as the focus of teaching and testing. She concludes that oral proficiency descriptors are often ambiguous, and are really instruments of norm-referencing, serving only to rank-order candidates.

A recent study of the limitations of multiple choice questions (MCQs) concludes that the construct validity of the format is open to question. In a study of Chinese learners' test-taking processes in a listening comprehension test Wu Yi'an (1998) used retrospection protocols to go into the mental processes lying behind candidate performance. The results suggest that the format favored more advanced learners and provided added difficulty for the less able. We will take up some of the details of this research later, when we look at the use of the MCQ test task in greater depth.

7.2.1.10 *Conclusions*

The concept of validity is more important than, but dependent on, that of reliability. In the literature, attitudes to validity vary, mainly according to the writer's intentions and perceived readership. Perhaps the most significant conclusion, which almost all would agree with, is that the purpose of the test defines the type of validity that needs most attention. In our study, as we show in the table below, we must consider all but one of the types we have defined. Only backwash validity escapes our purposes in that our test has no link with classroom teaching.

7.2.2 RELIABILITY

> *The consistency or stability of the measures from a test. The more reliable a test is, the less random error it contains. A test which contains systematic error, e.g. bias against a certain group, may be reliable, but not valid.* (A.L.T.E. 1998:160)

Reliability is concerned with the ways in which error can influence the scores of individual candidates. Three types of reliability are described in the literature: internal consistency, equivalence, and stability (Bachman 1990). Each of these looks at the reliability of the test from a different perspective. Internal consistency has to do with the differences in candidates' scores that arise from measurement errors produced by test tasks or item types. Equivalence is the relationship between different

Table 7— 4 Aspects of validity relevant to our study

Backwash	✗
Concurrent	✓
Construct	✓
Content	✓
Criterion-Related	✓
Empirical	✓
Face	✓
Non-Empirical	✓
Response	✓
Predictive	✓

forms of the same test. Moreover, test stability is a calculation of the sources of error that may arise over a period. All three types of reliability are calculated by the correlation between scores achieved on a specific test, and those coming from real or hypothetically parallel tests. The formulae used vary from type to type and, as with all correlational research, the figures they produce need to be interpreted with care.

The concept of reliability purports to demonstrate two things:

※ The value of a single language test as an instrument which can be used with a specific group of candidates on more than one occasion, or

※ The value of two or more parallel language tests which can be used with different groups of candidates on the same or different occasions.

In the context of large-scale testing such as that carried out by university authorities measuring school-leavers' abilities for entry to the university the concept is of great importance.

Whichever method is used, the question of the reliability of a test is fundamental to that of its construct validity: an unreliable test cannot be considered valid. However, the all-out pursuit of reliability can lead test designers to focus too narrowly on perfecting this aspect of test performance in an all-too-imperfect testing environment. Testing is much more a balancing act in which the different components of what Bachman & Palmer describe as Test usefulness (1996: 17-42) need to be held in balance. Reliability and construct validity are undoubtedly two essential components in this, but they cannot be considered the only two.

Testers have been seduced by the technical developments possible in the pursuit of greater reliability, at the expense of confirming validity (Spolsky 1995). The dangers of this tendency are particularly important in the handling of objective tests such as ours because they can lead to a disregard for the validity or validities of the test instrument.

Reliability, then, is concerned with the ways in which error can influence the score of the individual candidate. It can be defined in terms of the internal consistency of a test, the differences between test tasks and item types and of stability, sources of error that may occur over a period. The methods by which reliability is measured generally involve the calculation of correlations between real or hypothetically parallel tests. Internal consistency looks at the reliability of a single administration of one test. It is analyzed by studying differences in candidates' performance on different parts of the test.

If, as in our case, three different item types are used, how does this influence candidate performance? In studying equivalence, two aspects of item performance, THE FACILITY VALUE (FV) and the DISCRIMINATION INDEX (DI) can be used for comparison. In addition, by looking at the descriptive statistics of the overall performance of candidates on the test we can make certain deductions about the test as an instrument. In the absence of a second test, however, we need to adopt measures of internal consistency in order to establish reliability.

7.2.2.1 *Internal consistency*

> *A feature of a test, represented by the degree to which candidates' score on the individual items in a test are consistent with their total score.* (A.L.T.E. 1998:148)

Internal consistency is the reliability of the test itself, and of the scoring procedures used. It is an alternative to testing for stability, which we discuss below, and is calculated in circumstances in which a test-retest procedure — one in which the same test is administered twice — is not possible.

Aspects of the test method, such as item-type or target language, influence internal consistency. For example, if different item types appear in a test, how do they affect candidate performance? Internal consistency is an estimation of the reliability of a test when it is only used on one occasion, and we use different parts of the same test to provide the data we need. Two approaches to calculating internal consistency are described: split-half reliability estimates (Bachman 1990), and item variance reliability (Bachman 1990; Henning 1987). Both of these present certain advantages and disadvantages, which we will now describe in detail.

SPLIT-HALF ESTIMATES

If we divide the items in a test into two equivalent groups, we can compare candidates' scores for each half of the test as if they were two separate tests. We can then compare the mean and variance of each half, and the mean and variance of the whole test. Bachman (1990) proposes two methods of calculating split-half reliability: the SPEARMAN-BROWN, and the GUTTMAN formulae.

Whichever of these we adopt, we must make certain assumptions about the two halves of the test, and these inevitably affect our interpretation of the estimates we produce. In both cases, the formulae assume that the two halves of the test are independent of each other. That is, that the score achieved on one half cannot depend in any way on that achieved on items in the second half. Furthermore, the SPEARMAN-BROWN formula assumes that each half is equivalent to the other.

The apparently straightforward procedure of dividing a test into two equivalent and independent halves is clearly not so simple. Random selection, as in the so-called *odd-even* method - whereby *odd* items (numbers 1, 3, 5, etc.) and *even* items, (numbers 2, 4, 6, etc.) are separated - may be insufficient to guarantee independence and equivalence. Analysis of the two halves using the F-test, to compare variances, and the t-test, to compare means would, however, establish the validity of this approach.

In view of this difficulty, Bachman suggests that division "by design" may be more appropriate, and that...

> *[o]ne way of avoiding this would be to split the test into halves in every way possible, compute the reliability coefficients based on these different splits, and then find the average of these coefficients.* (175)

Although he does then point out the difficulties in terms of the sheer volume of calculations this entails.

One further complication of the split-half method is that by dividing the test in two, the total number of items in each half is reduced. As reliability improves when the number of test items is increased, this puts further doubt on any interpretation of the split-half estimates. To overcome this, a variation of the SPEARMAN-BROWN formula corrects for a small number of items (Bachman 1990; Henning 1987) The GUTTMAN SPLIT-HALF estimate does not require this type of correction.

ITEM VARIANCE RELIABILITY ESTIMATES

As an alternative to split-half estimates internal consistency can also be calculated by means of the KUDER-RICHARDSON RELIABILITY COEFFICIENTS, known as KR-20 and KR-21. These formulae calculate reliability from the variance of individual items (Bachman 1990; Henning 1987). However, they assume that all items are of equal difficulty, which may not be the case.

The assumptions underlying the use of the three formulae and the consequences of violating these assumptions are described in the table that follows. Whichever formula we use, the consequences of violating the assumption of independence and/or of equivalence is the same.

We define equivalence as the degree of measurement error derived from using different forms of the same test. We measure this when we correlate the scores achieved by a single group of candidates on two, or more, tests. Equivalence is calculated with the PEARSON PRODUCT-MOMENT COEFFICIENT.

We find more evidence of equivalence, or of its absence, if we study two aspects of item performance: FACILITY VALUE (FV) and DISCRIMINATION (DI). These values, which we will describe in more detail shortly, indicate how candidates respond to individual items. To compare different forms of the same test we can compare these values, and their means, along with other descriptive statistics.

Table 7 – 5 Assumptions for internal consistency reliability estimates, and effects of violating assumptions (Bachman 1990:178)

		Assumption		*Effect if assumption is violated*	
Estimate	*Coefficient type*	*Equi-valence*	*Indepen-dence*	*Equi-valence*	*Indepen-dence*
Spearman-Brown	Split-half	Yes	yes	Under-estimate	Over-Estimate
Guttman	Split-half	No	yes	--	Over-Estimate
Kuder-Richardson	Item variance	Yes	yes	Under-estimate	Over-Estimate

7.2.2.2 *Stability*

An aspect of reliability where the estimate is based on the test-retest approach. It relates to how stable test scores are over time. (A.L.T.E. 1998:164)

Stability, also known as test-retest reliability, answers the question "Do candidates perform differently when taking the same test for a second time?" However, it is a concept that is difficult to measure in most contexts, and one that we are not able to gauge.

The formula we have presented earlier is based on one administration of the test and calculates reliability from the responses by the candidates who sat the paper. It is designed to simulate the test re-test situation, in which a single group of candidates would take the same examination on two separate occasions.

7.2.2.3 *Spearman-Brown Prophecy Formula*

The test writers' chances of achieving high levels of reliability are enhance by the application of the SPEARMAN-BROWN PROPHECY FORMULA. In any given test, the level of reliability achieved may be lower than that targeted. With this data, we can calculate the improvements in reliability that would be achieved by adding more items to the same test. In this way, it is possible to report on the improvements needed in a highly specific manner. If a test is being trialled, this information can enhance the revision process.

7.2.2.4 *Standard error of measurement (S_e)*

Having calculated the coefficient of reliability for a test, it is then possible to produce the STANDARD ERROR OF MEASUREMENT. This figure indicates how close the observed score of candidates is to their true score, and as such is a means of gauging the overall accuracy of the test results produced. Crocker and Algina (1986) describe this process, but they offer no suggestion as to what might be considered an acceptable level of error.

7.2.3 *DISCUSSION: RELIABILITY VS. VALIDITY*

Classical Test Theory (CTT) defines these two concepts — reliability and validity — as essential to test analysis. Bachman (1990:161-62) distinguishes between them through questions. Of reliability, he asks:

> *How much of an individual's test performance is due to measurement error, or to factors other than the language ability we want to measure?*

And of validity:

> *How much of an individual's test performance is due to the language abilities we want to measure?*

The precise nature of the relationship between reliability is less clear nowadays. Writers (e.g. Heaton 1988) had generally accepted the view that reliability was essential to establishing validity, and dedicated much time and energy to the development of increasingly reliable tests. Spolsky (1995) describes in detail the manner in which testers had been seduced by the technical developments possible in the pursuit of greater reliability, which he considers to have been at the expense of confirming validity.

Nowadays, the nature of the reliability-validity relationship is said to be less clear, and it is considered a function of the purpose of the specific test under study, rather than an immovable concept. Bachman (1990) and Spolsky (1995) both stress the balance of the relationship pertaining between the two concepts, and emphasize the relatively greater importance of validity.

From our point of view, this means that while it is obviously desirable for us to achieve the highest levels of validity and reliability, we have to accept that this is not necessarily going to be possible. Throughout we must be keep our sights set on these targets, but we cannot afford to lose sight of the realities of carrying out large-scale testing exercises. We will have to establish target figures for the split-half reliability and the standard error of measurement, in order to gauge the reliability of our tests, and for the correlation coefficients deriving from our validation of test performance by comparison with other instruments.

7.2.4 *PRACTICALITY*

The practicality of testing is

> *[p]erhaps the most important quality of any test...* (Baxter 1997:27).

Also known as Test Efficiency (Weir 1983, 1988) the term is used — but not defined — to cover all of the real-world aspects of small and large scale testing which can easily call into question the efforts and best intentions of test designers, writers, administrators, takers, and users. Weir (1980) comments on the question of time as being one of the most important aspects of test efficiency. To this, he later (1983:76) adds three aspects of practicality under the general heading of "economy". He suggests tests should always be designed with these in mind:

- Ease of administration
- Ease of scoring
- Ease of interpreting.

It is his conclusion that in Communicative testing there is always something of a balancing act between considerations of validity, reliability, and practicality. This point is reinforced by Alcaraz Varó and Ramón y Denia (1980:22) who add a further condition with regard to classroom testing, namely that the test should be viable in terms of the teacher's time and of the students' time. It should also be practical in terms of the materials and resources that it requires. Furthermore, they underline the point that however valid and reliable a test may be, it is of no value whatsoever if it is impractical.

In writing primarily for the EFL classroom teacher, Baxter (1997:27-28) presents an idealized checklist, which we reproduce in adapted form below, setting out what are essentially the same components of practicality in a much more detailed form.

We have edited out some elements that are not directly relevant to our study, but feel that the general lines of Baxter's checklist are eminently applicable to our situation.

In the first instance, the questions arising from the limitations of time on the testing process were considerable. Throughout his table, Baxter uses the plural form "teachers". While it is highly

Table 7— 6 "What is practicality?" (Adapted from Baxter 1997:27-28)

Time will be needed for:
Teachers designing the test Teachers analyzing the results (e.g. how successful are the distractors) Trialling it on sample groups Teachers marking the papers Students doing the test
Writing a test which is valid and reliable requires:
Teachers who are experienced in test-writing Teachers who are expert statisticians Teachers to attend pre-marking standardization sessions Teachers to mark the tests
Space and equipment
Students need to be sitting where they can't copy (especially in multiple-choice tests) They may need different tables (e.g. one desk per person) They may need calculators or computers to record and analyze the results
Money for:
Extra staff Extra space Extra equipment (However, this money is probably not available.)

desirable to have teams of teachers involved in the process of test design and writing, the practicality of finding any number of individuals who fit the description Baxter presents ("Writing a test which is valid and reliable...") is highly idealized.

Secondly, we believe that there is a culture-based perception within the Spanish university system, which impedes the provision of adequate time for test analysis before the publication of results. This "blockage" is the concept of *libertad de cátedra*, and the assumption of independence of authority conferred on the members of examination boards.

Examination boards are generally made up of interested parties, but they are often put together using political, rather than academic criteria. It has been our experience that, sadly, the membership of boards represents "forces to be reckoned with" rather than Subject Matter Experts, or specialists in testing. This can lead to the accidental or even intentional misuse of examination scores. It is not the place of this research study to go further into these questions. However, in the light of the increasing concern within the world of testing for ever higher levels of ethical and social standards, achieved primarily through greater transparency, we feel that this aspect of test practicality must be mentioned.

In the context we worked in, one significant difficulty lay in the timing of the actual tests themselves. These, due to the requirements of the university admission process, had to take place in late July, after the main university entrance examinations. In addition to the high temperatures that teachers and candidates would have to endure on the days of the test, and in the following period of marking and decision-making, this is a period during which university teachers traditionally are free of timed work commitments. Consequently, many teachers consider themselves free to pursue their research interests, and many are involved in summer schools and doctoral courses away from their universities.

Our test focussed on reading skills both in LA and LB. The test was designed to ascertain whether candidates demonstrated an aptitude for translation sufficient to gain entry to a first degree course in Translation and Interpreting. However, the test in itself was not the only instrument by which they were measured, and it was not decisive in the entry process. For a number of considerations, which are outside the scope of this study, candidates were not rank ordered in terms of the scores they obtained in the entry test, but were simply divided on a Pass/fail boundary. In practice, candidates' entry test scores were scaled by a factor derived from their University entrance examination scores and ranked in order of this combined score. This meant that candidates with low scores in the test of aptitude, and high University entrance examination scores may be accepted to the course in preference to others with very high scores in Aptitude.

7.3 Facility value

> *The proportion of correct responses to an item, expressed on a scale of 0 to 1. (A.L.T.E. 1998:145)*

The Facility value (FV) of an item, also known as the Facility index, or *p*-value, is calculated be dividing the number of "correct" responses by the total number of subjects in the population (Heaton 1988:178-9).

In the design of a test, the decision may be taken to introduce a range of items that have different FV scores, or to target all items at the same FV score. In our case, we decided to vary the FV scores for the three sections of each sub-test. The first items were intentionally easier than those in the later sections because we wanted to "ease" candidates into the examination, as Henning suggests (1987:500). Later sections were correspondingly more demanding.

FV values ranging between 0.33 y 0.67 might be considered the ideal (Tuckman 1978, quoted by Henning 1987:50). However, as Heaton (1989) points out many test writers would be happy to accept a slightly wider range: between 0.3 and 0.7, say. In our context, we expect higher FV scores for the Spanish sub-tests, than for those in English.

7.4 Discrimination

> *The power of an item to discriminate between weaker and stronger candidates. (A.L.T.E. 1998:143)*

In CTT uses two means of calculating the discrimination of an item. One involves the correlation of the item score with a criterion such as the total test score or an external measure. The other involves comparing the difference in difficulty for weaker and stronger students. This latter procedure gives rise to a score known as the DISCRIMINATION INDEX (DI), and it is figure that we use in our analysis.

7.5 An alternative to CTT?: Item response theory

Item response theory (IRT) is described as one of the most powerful approaches to test reliability modern measurement theory has developed (Baker 1997; Muñiz Fernández 1990). It aims to provide test users with an estimate of candidate performance based on both the difficulty of each specific item and on the individual candidate's level of ability.

IRT makes certain assumptions. Firstly, there is a general assumption of 'unidimensionality' This states that each item is testing one and only one ability trait. Secondly, it makes specific assumptions about the relationship that exists between the learner's level of ability and the item that can be plotted on an Item characteristic curve (ICC). This relationship is assumed to be non-linear, and it connects the level of ability, measured from 0 to 1, with the probability of correctly responding to the item, measured in terms of units of Standard deviation from -4.0 to +4.0. Each ICC can be studied for one, two, or three parameters. These are known as the Discrimination, Difficulty, and Pseudo-chance parameters. Two parameter models discount the guess factor represented by the third of these, and one parameter versions - the most widely used of which is the Rasch model - ignore this and also assume that discrimination is the same for all items.

The conclusions drawn from applying the three-parameter IRT model to a set of test scores divide in two. The Item information function (IIF) illustrates the value of any one specific item at a specific level of ability. It can do this as the ICC shows the levels of discrimination and difficulty for each item at each level of ability, as designated by measures of Standard deviation. The sum of the IIFs is known as the Test information function. This offers the global appreciation of a test, and is IRT's equivalent of CTT's measures of reliability and standard error. The most important advantages are that the information provided by IRT methods is independent of sampling, and is much more precise offering estimates at each ability level. The disadvantage lies in the costliness of the computer based, analytical processes which render IRT low in cost-effectiveness, and therefore of little practicality for our study.

7.6 Conclusions

In this chapter we have tried to describe in some detail the main concepts that make up Classical Test Theory in order to describe the terrain in which we are going to carry out the practical aspects of our study. These concepts amount to the criteria by which we will judge the results of our research. In addition, we have surveyed some of the analytical principles that are widely applied within CTT in order to obtain empirical data that can be manipulated and discussed.

Part II

Research design and methodological considerations

We are fully aware that in research of this nature, "trade off" must exist between demands of validity, reliability, and practicality. We will probably have to accept that our tests are unlikely to achieve our targets for all three of these complex criteria. Given the demands of the real-world context in which we are working it seems likely that practicality will have to take precedence over validity and reliability.

8 APPROACHES TO RESEARCH DESIGN: CORRELATIONAL RESEARCH IN THE VALIDATION OF APTITUDE TESTS

8 APPROACHES TO RESEARCH DESIGN: CORRELATIONAL RESEARCH IN THE VALIDATION OF APTITUDE TESTS

8.1 OBJECTIVES
8.2 RESEARCH DESIGN
8.2.1 *Correlational research*
8.3 CARROLL'S SIX CONDITIONS TO TEST THE VALIDITY OF A MEASURE OF APTITUDE
8.4 OVERALL RESEARCH QUESTION
8.4.1 *Specific research questions*
8.4.2 *Hypotheses*
8.4.3 *The null hypothesis*
8.5 TEST DESIGN
8.5.1 *Introduction*
8.5.2 *Language test construction and evaluation*
8.6 STEPS IN THE TEST-WRITING PROCESS
8.6.1 *Writing a test specification*
8.6.1.1 The entry test's purpose
8.6.1.2 The candidate profile
8.6.1.3 Specification of test content
The first draft specification
8.6.2 *Finding texts to meet the specification criteria*
8.6.3 *Item writing*
8.6.4 *Moderation of items*
8.6.5 *Revision and formatting*
8.6.6 *Analysis of candidates' scripts*
8.6.7 *Administration of tests*
8.6.8 *Marking of scripts*
8.6.9 *Computer analysis of results*
8.6.10 *Statistical analysis: Classical Test Theory (CTT)*
8.6.10.1 Internal consistency
8.6.11 *Reporting scores and setting pass marks*
8.7 SAMPLING
8.7.1 *Population*
8.7.2 *Sample size*
8.7.3 *Sampling error*
8.7.4 *Sources of error*
8.8 VALIDATION AND TEST ANALYSIS
8.8.1 *Oxford Placement Test*
8.8.1.1 Analysis of the OPT Listening/Reading sub-test
8.9 CONCLUSIONS

IN THIS CHAPTER we will describe both in general and specific terms, the design considerations that went into the planning and execution of this project. Our objectives are

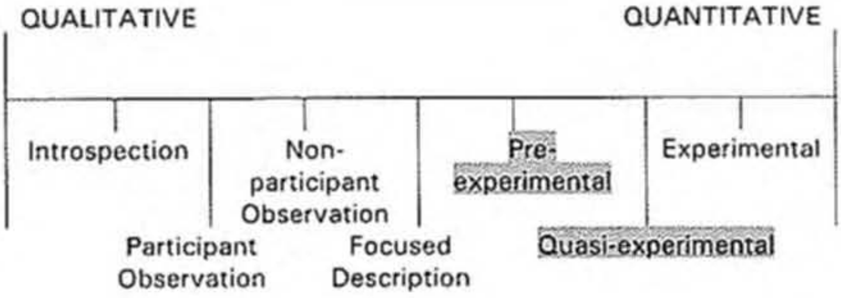
- To place our study within the context of research paradigms current in the social sciences and humanities
- To establish an ordered test design, development, and refinement process
- To detail the stages of that process with specific reference to our test
- To set the parameters and statistical targets by which we intend to judge the performance of our sub-tests

8.1 Research design

The dichotomy established between the normative and interpretative research paradigms after Cohen & Mannion (1989) is one which is called into question by Black (1993:1-23) who stresses the essentially complementary nature of the relationship between the two models. While a normative approach sees human activities as essentially belonging to the field of rule-governed behavior, and derives its methodology from the natural sciences, the interpretative paradigm emphasizes a concern for the individual. These, subsequently, can been seen to become the quantitative and qualitative approaches to research. In the case of our study, which is more of the former type, dealing as it does with such large populations, it is hoped that the volume of material has not overwhelmed an essentially qualitative concern. The purposes of the test in aptitude for entry to first degree studies in Translation and Interpreting has the overriding aim of keeping to a minimum the level of failure and consequent dropout rate. It seeks to do this in an area of studies that is highly demanding in the light of the experience of the Diploma course, which the degree program replaces.

The opposition between the qualitative and quantitative research paradigms has been minimized by a number of authors (Cohen and Mannion 1989; Seliger and Shohamy 1989; Larson-Freeman and Long 1991; Black 1993; Creswell 1994). Larson-Freeman and Long describe the relationship as that of two distinct positions on a continuum (1991:21), and their graphic presentation of this reduces the tension between approaches.

Table 8— 1 The qualitative-quantitative continuum of research methodologies (Larson-Freeman & Long 1991:15)



Methods from the quantitative end of this continuum are characterized by the degree of control the researcher is able to exercise over the variables under study, and the naturalness of the setting (Cohen and Mannion 1989; Larson-Freeman and Long 1991).

We believe that the present study will bridge the "pre-experimental" and "quasi-experimental" modes — shaded in the table above — in that it is in two parts. Firstly, there is a pre-experimental part involving the application of Classical Test Theory (CTT) to the test results. This analysis is an isolated process that produces a reliability coefficient to indicate the value of repeated use of the test. Secondly, there is a quasi-experimental part to the research, as the results we will study deal with the parallel administration of tests to different groups, and the study contains many elements of replicability.

Whichever label we use the limited value of our research as a measure of causality must not be ignored. The study we propose involves a correlational analysis to examine test validity, as defined by CTT. This can only indicate possible relationships, even though the quasi-experimental research design criteria for control and randomization will be more fully met in the detailed item analysis. Here, total control over the application of the test is possible, as is randomization in the selection of participants.

The second dichotomy in research design described in the literature is that which contrasts longitudinal and cross-sectional studies. Again, this study offers aspects of each. The longitudinal study is prospective in that it typically deals with the same respondents over a period of time, and purports to measure their behavior, and changes in that behavior. Our study of validation is typical in that recorded data is hypothesized to predict behavior, and over a period of two years, markers will be used to judge the accuracy of that prediction. From the point of view of the validation of the test, the variables that might influence these markers cannot be controlled: over a long period of time innumerable factors are likely to influence candidate performance. However, in our use of reading skills as one possible instrument to measure aptitude for translation, we accept the influence of these variables. We view aptitude as a trait which specific studies and training can develop in the individual: performance is therefore likely to be enhanced in those who possess a higher degree of aptitude. The study is also cross-sectional, in that the reliability coefficient and the item analysis are based on the specific administrations of the tests.

We believe that this study exemplifies the contention that "... the methodological design should be determined by the research question" (Larson-Freeman and Long 1991:14). However, this is modified in that the "naturally occurring setting" where our research takes place heavily influences aspects of the design, too.

8.1.1 *CORRELATIONAL RESEAR CH*

Correlational research is the study of relationships between factors arising in the complexity of human behaviour (Cohen and Mannion 1989). It is typically used in predictive studies and can never be considered to provide results indicating causal relationships. Three types of correlation exist: the simple correlation of two variables whether on a continuous or discrete scale; multiple correlation of more than two variables; and partial correlation, in which one or more of a number of factors can be eliminated. For the validation of the test mechanisms in this study, the PEARSON PRODUCT-MOMENT simple and multiple correlations are used common in studies of academic achievement (Bossé-Andrieu 1981). For other measures, the SPEARMAN-BROWN and GUTMAN split-half formulae, SPEARMAN-BROWN PROPHECY FORMULA (Crocker and Algina 1986), and KENDALL'S TAU$_b$, (Kendall 1975) formulae are applied (Silver 1997).

The predictive value of a correlational study can be seen through aspects of the coefficient of correlation once this has been calculated. These are the strength of the correlation, and the level of statistical significance.

Correlation coefficients are computed to give values between +1.0 and -1.0, with zero representing the absence of any correlation. Extreme values of either sign are highly unusual, and the strength of a correlation is interpreted without taking into account the sign. Correlations are grouped together and defined as follows:

Table 8—2 Defining the strength of correlation coefficients (Cohen and Mannion 1989:168-69)

0.20-0.35	Are of limited use.
0.35-0.65	Are of little use unless corroborated by more than one study.
0.65-0.85	Within this range correlations are "accurate enough for most purposes"
>0.85	Strong correlation

According to Cohen and Mannion, 0.40 is the absolute minimum for "crude" predictions, justified only if the evidence of multiple regression analysis supports them. The stronger the correlation, the more powerful it is as a predictive tool. However, our reading in the field of aptitude testing indicates that for practical purposes these parameters are modified to take into account the nature of the population being studied. Ehrman (1995) accepts 0.20 as the minimally significant correlation when dealing with a highly homogeneous sample.

A further point of importance is the level of significance established for calculating the correlation coefficient. Normally a 95% level, or $p = 0.005$, is considered essential, but this must be considered in relation to the size of the population, and the level of significance is best interpreted with reference to tables of significance such as those presented in Siegel and Castellan (1988).

8.2 Carroll's six conditions to test the validity of a measure of aptitude

As we have stated in Chapter 4, the correlational research design model that J.B. Carroll proposes (1981), and which we reproduce for a second time in the table below, seems the most appropriate basis from which to construct our study.

If we begin by looking at the different components of his outline we can see the extent to which using the design is practical in our circumstances.

All of the basic elements are included in the table. The two measures that are missing are, as Carroll points out, the two measures that would totally lack face validity. Namely, the administration of an achievement test B_{ix} at the time t_i of the entry test; and the administration of an aptitude test A_{fx}, at the time t_f of the final achievement test.

Table 8—3 Carroll's six conditions to test the validity of a measure of aptitude (1981:289)

Initial time T_i	*Period of learning on task x*	*Final time* t_f
Aptitude test A_{ix}		Aptitude test A_{fx}
(Condition 1: Reliable true score variance in A_{ix})	(Condition 4: No change from A_{ix} to A_{fx})	
(Condition 3: $r_{AixBix} = 0$)	(Condition 6: $r^2_{AixBfx} > 0$)	
Achievement test B_{ix}		Achievement test B_{fx}
(Condition 2: No reliable true score variance in B_{ix})	(Condition 5: Significant change in mean from B_{ix} to B_{fx})	

Our inability to apply these two elements of the research design quite logically renders most of the six conditions useless. Without an initial measure of achievement, Conditions 2, 3 and 5 cannot be proven. Moreover, without a final aptitude test, Condition 4 cannot be proven either.

Table 8—4 The essential components of Carroll's six conditions (1981) adapted to suit the circumstances of our test.

Carroll's theoretical components		*The realization of each component in or situation*
Measures of time	Initial time t_i	The date on which candidates take the entry test
	Period of learning on task x	Two years
	Final time t_f	The end of the first cycle of studies
Measures of aptitude	Aptitude test A_{ix}	The entry test
	Aptitude test A_{fx}	None administered
Measures of achievement	Achievement test B_{ix}	None administered
	Achievement test B_{fx}	General translation A-B General translation B-A examinations

Table 8—5 The realities of attempting to prove Carroll's six conditions (the shading indicates those conditions that *can* be met)

Condition	
1	Reliable true score variance in A_{ix}
2	No reliable true score variance in B_{ix} We assume this to be the case, but do not test for it.
3	$r_{A_{ix}B_{ix}} = 0$ This cannot be tested as there is no Achievement test B_{ix}
4	No change from A_{ix} to A_{fx} We assume this to be the case, but do not test for it.
5	Significant change in mean from B_{ix} to B_{fx} This cannot be tested either, as there is no Achievement test B_{ix}
6	$r^2_{A_{ix}B_{fx}} > 0$

As Carroll makes clear, we can only hope to prove Conditions 1 and 6. However, the concept on which Condition 1 depends is one that can only be demonstrated indirectly. The TRUE SCORE can only be estimated be calculating the STANDARD ERROR OF MEASUREMENT (S_e) of the observed score, and accepting this figure, plus or minus, as the range within which the true score lies. Consequently, Condition 1 cannot be proven empirically, but we can judge the value of the margin of error. Crocker and Algina (1986:146-152) describe the complexities of interpreting Standard error of measurement (S_e) values but give no indication as to a minimally acceptable level. Their argument is that test designers should always be aware of the margin of error involved, and that they should minimize this.

Condition 6, is a straightforward correlation of scores, and can be calculated by rank ordering candidates and applying the PEARSON PRODUCT-MOMENT CORRELATION or the KENDALL'S TAU$_a$ formulae. The question here is simply a matter of recording the statistically significant level of correlation, given the size of the individual sample.

8.3 Test design

8.3.1 *INTRODUCTION*

The essence of our research lies in the design and administration of the reading skills sub-tests that we have prepared, developed and refined in order to measure aptitude for translation.

In this part of the chapter, we move on to more specific aspects of the research. Here we report on our investigation of the practical aspects of test design derived from the theoretical input we have earlier described, and on their practical application. In doing so, we describe the research process as we intend to carry it out.

8.3.2 *LANGUAGE TEST CONSTR UCTION AND EVALUATION*

Although the question of test design has been partially described by many authors (e.g. Madsen 1983; Heaton 1988), it is only in more recent years that the entire procedure has become the focus of attention. The development of the testing industry documented by Spolsky (1995) has brought with it an interest in specifying the stages in what previously had tended, it would seem, to be a relatively intuitive process. The great commercial interest in the use of tests such as TOEFL, and the UCLES examinations, probably the most widely known testing organizations, has been part of the coming-of-age of the testing industry. With it, has come a realization that testing must be a more formally structured process.

In the following pages, we compare the work on test design of three of the most significant groups of test researchers. These are the work of Alderson, Bachman, and Brendan J. Carroll, and their respective co-authors (Carroll B.J. 1980; Carroll B.J. and Hall 1985; Alderson, Clapman and Wall 1995; Bachman and Palmer 1996). In addition, we will draw on a number of other writers who have contributed to this debate.

For different purposes, and from slightly different perspectives, all three teams of researchers offer a view of the overall test design process. While Alderson et al are less pragmatic in their focus, their work presenting as it does a questionnaire and the responses made by 12 UK-based examining boards, Bachman and Palmer, and Carroll and Hall, are much more structured.

Essentially, the test design process is divided into three stages (Table 8—7). These are closely linked to the stages posited by J.B. Carroll (1981), which we have adopted as our research design model. Although the elements included at each stage do not always coincide, there is clear agreement among the authors as to the relative importance of each component, and little disagreement about the order of each stage in the overall process.

Table 8—6 The test design process

	Alderson et al (1995)	*Bachman and Palmer (1996)*	*Carroll B.J. (1980); Carroll B.J. and Hall (1985)*
1	Specification	Design	Design
	Item-writing and moderation	Describing	Description of participants
		Identifying	Analysis of communicative needs
		Selecting	Specification of test content
		Defining Developing Allocating Managing	
2	Pre-testing and analysis	Operationalization	Development
	Examiner and administrator training	Selecting	Realization of tests
	Reliability Reporting scores and setting pass marks	Specifying	Trial application
		Writing Administration Administering	
		Collecting feedback	
	Validation	Analyzing	Validation and test analysis
3			Operation Full-scale application Operational use
	Post-test reports Development and improvement		
		Archiving	Revision of test system
	Standards		

The differences that we encounter between the three approaches can largely be put down to their different intended readers, rather than to any fundamental differences of view.

8.4 Steps in the test-w riting process

We will next look at the process we planned to go through in writing our tests and in their subsequent analysis and revision. In doing so we draw on the input of all three of the approaches outlined above, among others.

8.4.1 *WRITING A TEST SPECIFI CATION*

To begin with, we will analyze the following summary of a typical test specification written about the University of Michigan English Language Institute Examination for the Certificate of Competence in English (ECCE). Spaan (1994) reports on the background to this examination and on the exam specification. The ECCE is aimed at intermediate ESL candidates, aged between 12 and 16 years or adults. The test is of communicative language use and is specifically not "academically-oriented". It is set at a level similar to the Cambridge First Certificate examination (TOEFL 425-525 points). The full examination comprises two elements: oracy and literacy, each of which is subdivided into a series of tests. The literacy component covers grammar, vocabulary, and reading in 80-120 items in total, taken over a period of 65 to 90 minutes.

The reading section involves "about" four readings of different genres of text, and the texts chosen are "ideationally accessible" (Spaan 1994:29), so as to ensure that candidates neither require any prior knowledge of the subject matter, nor are advantaged by any knowledge they may have. There are between 20 and 25 questions, and these are divided into a set of tasks which entail the use of a range of reading strategies.

Task 1 is based on a text specified as being a "traditional narrative". Between five and eight items of a general comprehension type are set. These are designed to elicit the main idea of the text, to distinguish this from the supporting details, and to analyze the relationship between the two. We interpret these to be equivalent to the range of skills and sub-skills defined as "Distinguishing the main idea from supporting details" (Munby 1978:126-131 Skill 39). The tasks involve a set of questions followed by one or more texts. There are either two or three shorter paragraphs, of between 50 and 80 words, drawn from advertisements, postcards, notices, or similar genres; or one longer text, of between 250 and 300 words, which might be a description or a set of instructions. The former text type is used to test candidates' abilities to skim and compare texts, with items focussing on context, and/or the relationship between parts of the text.

In order to interpret the test more easily, we equate these skills with those defined by Munby (1978:126-131) as

> *Skimming to obtain the gist of a text (Skill 45.1)*

or

> *Skimming to obtain a general impression of the text (Skill 45.2)*

and

> *Understanding relations between parts of text through grammatical cohesion devices of comparison (Skill 32.2).*

The longer text is used to test for "Scanning to locate specifically required information (Skill 46); "Understanding explicitly stated information" (Skill 20)

The third task type set in this examination is a 10 to 15 minute activity based on the carrying out of a written production task through the comprehension of a reading passage. The passage might take the form of a set of instructions, and by following these, candidates would complete a form. The data to be read would be contained in one or two short paragraphs, similar to those described in task 2, and the instructions would require short answers. The skills involved would be those of

> *Understanding explicitly stated information" (Skill 20)*

in order to follow directions, and understand description. This third task type is reported by Spaan as being difficult for candidates in trials, and consequently a question mark hangs over its continued use.

The ECCE grammar component uses questions of the multiple-choice type. It consists of 30 to 40 items based on single sentences, or on two or three sentences per item, where these are used to contextualize the target item. The vocabulary component has a similar format, but the items are of the multiple-choice gap-filling type, again based on context and appropriacy. The specification is precise in that it excludes the use of highly idiomatic or specialized vocabulary.

In a doctoral thesis reported by Davidson (1994:80), Dongwan Cho, at the University of Illinois at Urbana-Champaign, has studied "The Effect of Specificity of Language Test Specifications on Item Construction". His work has produced the apparently contradictory conclusion that while specialists found tests written according to less specific guidelines were more satisfactory, a combination of statistical analysis using

Classical Test Theory, Item Response Theory, and distractor analysis, indicated that more specific guidelines led to the production of "better" items. While he concludes that the communication between test specification writer and item writer(s) should lead to the most successful writing of items, he seems inclined to favor a less specific base for item writing. The final word, though, he leaves to further quantitative and qualitative research.

Our reading of this research leads us to conclude that, in practice, there is a significant lack of precision in the preparation of examination papers, and much of this has to do with the degree of detail with which the specification is written, and the item writers' interpretation of this.

Given the unique characteristics of the context in which we work, we have chosen those elements of test specification that we believe are the most important for our purposes. We will now discuss these in the light of the considerations found in the literature, the constraints of our situation, and with an awareness of the lack of precision in test writing to which we have just referred.

8.4.1.1 *The entry test's purpose*

The purpose of the entry test is

☛ To select, from among the potential candidates, those with the highest level of aptitude for translation and interpreting for entry to the first degree program.

The reading skills sub-tests, which are the object of this research, are one component of the entry test, and are specifically aimed at the selecting candidates for their potential ability as translators.

The entry test is a classic example of a political instrument. It is one of society's "door-keepers", aimed at allowing "suitable" candidates to enter, or at keeping "unsuitable" candidates out — depending on one's standpoint. The questions of ethical standards, raised by Alderson et al (1995:235-260) are of paramount importance when we look at the test and at its purpose. The establishment of agreed principles by which the entry test can be judged by all of those involved — designers, writers, markers, administrators, candidates, potential employers, in short, society at large — is a fundamental consideration in all testing. But, sadly, it is a fundamental consideration that has yet to be accepted on any significant scale. Alderson et al report that there is much "work in progress" (259). And it is our experience that even at the core of the examination "industry" there is still very much to be done.

One important aspect of the role of this entry test is that which we discuss in more detail below when we analyze the constraints on reporting and using the test scores produced by the entry test.

8.4.1.2 *The candidate profile*

One of the most important considerations Alderson et al mention in their description of the specifications of tests is the relationship existing between the specification itself and the syllabus to which it is linked. This is also a basic consideration of Carroll's work, in that he derives his test design from Munby's work on the communicative syllabus (1978), all of which was a part of the attempts to ensure that language teaching grew closer to the learner.

In the case of our test, which is for entry to a course of study, there is no syllabus as such to take into account. However, the considerations as to the profile of candidate must draw on two elements of supposed knowledge as a starting point. These are the supposed profile of the student who has successfully completed secondary schooling in Spain, and who has achieved entry to the university, and the supposed knowledge of the first year undergraduate student in Translation and Interpreting.

Our profile of the candidate, therefore, is two directional. In the first case, we can "look back" as it were, to the learning achieved; in the second case, we can "look forward" to the learning projected. The legal background of the entry test, which we will discuss in more detail below, specifically forbade testing knowledge or skills that would be taught as part of the studies in Translation and Interpreting. However, the language skills candidates could be supposed to have developed in secondary schooling would not be taught again, as if unknown, but would certainly be "revisited" as part of the cyclical development of language typical in FL course and syllabus design.

Consequently, our first conclusion in terms of the candidate profile was that candidates would have attained a significant base in communicative language skills. These skills, we believe, would almost certainly be of the type originally defined by Munby (1978), as these form the general base of Communicative language teaching and learning widespread throughout Spain, and indeed the world.

Our second conclusion based on the experience of the Diploma course, which we described in an earlier chapter, was that the minimum level of ability that we could expect to find among candidates would be at, or below, that of the Cambridge First Certificate examination. This assumption lacks detailed supporting evidence, and it is one element of the test and research procedure that we have to take seriously. We cannot afford to introduce a test that is subsequently going to demonstrate that it has brought about a lowering of standards.

8.4.1.3 *Specification of test content*

While the purpose of our test is clearly not communicative in the way in which this is defined in the literature (Bachman 1990) and applied by Carroll (1980) for example, it remains an evidently functional test. The overall purpose of the test has been described above. The focus of the test is largely delimited by the legal specification published by the Ministry of Education. The essential elements in this are that:

- Performance should not depend on candidates' prior knowledge or maturity

- The test should be in line with the requirements of those centers forming part of the Permanent International Conference of University Institutions involved in the training of translators and interpreters (CIUTI)

- The test should measure candidates' control over their aptitude for linguistic transfer:

> *.... unas pruebas que, sin pretender valorar conocimientos previos ni siquiera madurez, puedan detectar de forma global las aptitudes del candidato para ser traductor e intérprete ... pruebas de evaluación en orden al control de aptitudes para la traslación lingüística demostrada en base a uno de los idiomas extranjeros conocidos ...(B.O.E. Nº 228:32181)*

Our interpretation of the first of these constraints has clear consequences in respect of the selection of test materials and mechanisms: the test cannot include any material prior knowledge of which would constitute an advantage to some of the candidates. In Spaan's terms, it must be "ideationally accessible". Similarly, the test mechanisms need to be of a type candidates can be expected to have previously encountered, and these should be clearly explained in such a way as to ensure they have no influence on candidate performance.

The question of the status of those criteria expected by the CIUTI is a different matter. The debate is, in fact, circuitous as the criteria accepted by the CIUTI are those designed and used on an individual basis by its members. The University of Granada applied for membership of this organization during the final years of the diploma course and was rejected. One of the grounds for this was the fact that a specific aptitude test for interpreting was not at that time part of the entrance procedures.

The third stipulation was the most significant in terms of the specification of contents and mechanisms. However, it opened a series of questions. How can 'linguistic transfer' be measured without requiring candidates to translate? We have decided that, in view of our reading of translation theory, we can adopt the use of reading skills as a measure of aptitude for translation as these constitute a prerequisite of the utmost importance.

THE FIRST DRAFT SPECIFICATION

The first draft test specification, in Table 8—8, was presented in Spring 1993. This specification outlines the use of three sub-tests:

❶ LA reading skills

❷ LB reading skills

❸ LB listening skills

The draft specifies that the reading skills tests should be of the multiple choice type, and further specifies the texts on which each of the sub-tests should be based. These specifications are in terms of text type and word length.

Careful examination of the draft (Table 8— 7) will in fact reveal one important contradiction in its content, in that the use of multiple choice tasks invalidates the inclusion of the skills listed at numbers 10 and 12, as these are not easily susceptible to testing in this manner. Both transcoding and note-taking are skills that are best examined in more open tasks, although this does not mean that they cannot be marked objectively.

One of the aspects of the administration of the test that appears in the specification needs clarifying in the light of the research in progress. Although the specification is that the sub-tests in Language A should be eliminatory, for ease of administration all candidates would take both the LA and the Language B sub-tests in consecutive sessions. In this manner, we would have the relevant data for all candidates, although the reporting of scores would follow the procedure indicated.

Table 8—7 The first draft specification for the entry test (Faculty of Translation and Interpreting 1993)

Licenciatura en Traducción e Interpretación
PROPUESTA DE PRUEBA DE APTITUD
1 Especificaciones
1.1 Formato

La prueba constará de tres ejercicios obligatorios:
1° Prueba de habilidades lectivas en lengua A (Español)
2° Prueba de habilidades lectivas en lengua B (Alemán, francés o inglés)
3° Prueba de habilidades auditivas en lengua B

1.2 Puntuación

El objetivo del sistema de puntuación es el de adjudicar las plazas vacantes a los alumnos más adeptos de entre los que se presentan.

El primer ejercicio tendrá carácter eliminatorio. El segundo y el tercer ejercicio se evaluarán conjuntamente.

1.3 Descripción de los ejercicios
1.3.1 1er ejercicio

Los candidatos deberán responder a una serie de preguntas de elección múltiple sobre un texto periodístico de actualidad, de una extensión aproximada de seiscientas palabras, con el fin de determinar su capacidad de usar las siguientes habilidades lectivas:

1. Deducir el significado y uso de unidades léxicas mediante la comprensión de la morfología y el contexto
2. Comprender las relaciones establecidas entre los componentes de una frase
3. Comprender las relaciones establecidas entre distintos componentes del texto que reflejan los nexos gramaticales
4. Comprender las relaciones establecidas entre distintos componentes del texto mediante el reconocimiento de indicadores de discurso
5. Comprender la función comunicativa de la frase según se empleen, o no, indicadores de discurso
6. Comprender las relaciones conceptuales, p.ej. de causa y efecto
7. Comprender ideas explícitas
8. Comprender ideas implícitas
9. Separar el contenido esencial de los no esenciales, distinguir la idea principal de los detalles que la ilustran
10. Transferir información desde un medio a otro, p.ej. desde un texto a un gráfico
11. Examinar el texto rápidamente para obtener información específica
12. Tomar apuntes. Reducir el texto mediante la eliminación del contenido redundante o irrelevante

1.3.2 2° ejercicio

Los candidatos deberán responder a una serie de preguntas de elección múltiple sobre un texto periodístico de actualidad - pero no de opinión - y de una extensión aproximada de cuatrocientas palabras, con el fin de determinar su capacidad de usar sus habilidades lectivas en la lengua B.

1.3.3 3er ejercicio

Los textos a utilizar serán monólogos grabados de medios de comunicación públicos - radio o televisión. Serán de interés general, sin ser de opinión, ni semi-especializados. El total de la grabación durará unos siete u ocho minutos. Los acentos reflejarán variedades de uso común a nivel internacional: p. ej. Inglés británico e inglés americano.
Se formularán las preguntas con el fin de averiguar la capacidad de los candidatos para usar sus habilidades auditivas.
2 Tribunal
El presidente del tribunal será el Director de la EUTI, o la persona en que él delegue la responsabilidad. Además del presidente, el tribunal estará compuesto por un secretario/coordinador de las pruebas de aptitud y nueve vocales, en representación de las lenguas B.

8.4.2 *FINDING TEXTS TO MEET THE SPECIFICATION CRITERIA*

Finding texts is a notoriously difficult process, and as Alderson et al (43) point out, one of the advantages of writing examinations on a professional scale is that the test writer tends to collect a supply of potential texts from which to choose. In the context in which we began our work, this was not possible, although later sub-tests clearly benefited from the extended timescale involved, as the authors were able to gather material over a much longer period.

In fact, before writing the reading skills sub-tests es01.1 and in01.1 texts that might serve as test material were collected over a very brief period. Source materials were taken from the national Spanish press, from the "quality" British English press, and from weekly journals of both British and American English, such as *The European*, *The Economist*, *Time*, and *Newsweek*. By cross-checking these with the criteria in the draft exam specification, we identified possible texts, and included among these, texts that might be suitable with or without editing.

8.4.3 *ITEM WRITING*

In order to write individual items we selected points of interest in the texts after reading and re-reading them to gain a general impression. In this way we tried to identify elements of the texts that were, potentially, suitable as the focus of test tasks. The texts chosen were judged subjectively to be both of a length and level of difficulty that candidates could be expected to have encountered previously in their secondary LB studies. At that stage, we did not have the technical means to calculate the readability levels of the texts. Measures of text readability, such as the Flesch or Flesch-Kincaid scales have only recently been included in standard word-processing packages such as Microsoft®Word 97 (Microsoft Corporation 1983-1996). Moreover, they are not necessarily as sound as they might seem (Sydes and Hartley 1997) even when the software operates efficiently.

Following the decision to work on a specific text, this was studied to identify elements of cohesion and coherence; topic sentences, and keywords were identified; and the overall structure of the text drawn out. In particular, elements of grammatical or lexical cohesion were identified in order to relate these to the target skills in the test specification.

By reference to the list of skills in the first draft specification (based on Munby 1978) possible sources of test items within the text were signalled. At this point, it was important to ensure a representative coverage of the skills in the specification although it was clear that no text would necessarily allow us to write items that covered all of these.

This process led to the drafting of test items according to the list of task types included in the specification. Writing test items is considered much more an "art" than a "science", and experts regard the process of item writing as long and difficult, but creative and satisfying (Carroll, Brendan J. Personal communication 1993). It is a slow process, and a professional item writer considers that a "production rate" of three items per full working day, is highly satisfactory (Hootjes, G. Personal communication 1994).

Writing usually began with identifying the key answer. To this, we would then add an appropriate question stem. Finally, we would look for the required distracters to complete the item, according to the task type in question. The source of distractors was often the co-text around the key response, or other parts of the text that dealt with the same or similar elements. In other instances, distractors were found by working "around" the lexis or syntax of the key response itself. As it was not always possible to find all of the distractors immediately, some items were left incomplete at this stage, with a full set of options being drafted later.

8.4.4 *MODERATION OF ITEMS: CONSTRUCT VALIDATION*

One of the essential aspects of item writing stressed by both researchers (Alderson et al 1995; Carroll, Brendan J. Personal communication 1993) and Examination boards (Evans Personal communication 1993) is the collaborative nature of the process. In our case, one or more colleagues reviewed the first draft of each test paper. Their task was twofold: firstly, they were asked to apply their own knowledge of testing and testing procedures to analyse the items and point out any potential difficulties. Secondly, they examined the construct validity of items by comparing them with the taxonomy of reading skills on which they were based, and introspecting on the validity of the items as appropriate tests of these skills.

8.4.5 *Revision and formatting*

The finalised papers were prepared for use. This work involved the standardised word processing of the tests to ensure quality and clarity of presentation; the formatting of question papers and answer sheets; the preparation of instructions for candidates; and the adaptation of the texts for photocopying, which involved pasting on letters to identify the paragraphs.

8.4.6 *Analysis of candidates' scripts*

Prior to administering the tests, the markschemes containing the key responses were shown to some colleagues for them to check the responses. They were asked to carry out the tests as if they were candidates and to try to establish whether the process of answering the items gave rise to any anomalies. Minimal revisions were made in the light of this, and these principally involved altering the order of the distracters to ensure the number of correct responses for each of the alternative letters (A, B, C, D, or E) was more or less the same. A preponderance of any one letter might induce candidates to commit errors by changing their responses so that their answer sheets offered a full range of options. This kind of "over-compensation" is not particularly common, but it does occur when candidates have dedicated time to preparing for objectively marked test tasks of the type we were using (Evans Personal communication 1994).

8.4.7 *Administration of tests*

The University entry process in itself has a significant impact on the timing of the administration of the entry tests. Due to the nature of the University entrance procedure, applicants in the June sessions in 1993 and 1994 were first required to complete their general University entrance tests, *selectividad*. Secondly, they had to make a preliminary application to study at the University of Granada in the forthcoming academic year. Finally, they needed to make a specific application to take the entry test in order to gain access to this program. Each of these steps was accompanied by a fixed timescale, which had to be adhered to. Consequently, the test could not be held before the third week in July. Clearly, this meant considerable time pressure on those University lecturers who were to make up the

Examining board responsible for the Test as well as the inevitable pressures on the applicants.

8.4.8 *MARKING OF SCRIPTS*

Members of the Examination board were to mark the scripts using plastic overlays. Their marking was to be double-checked by other members of the board to ensure neither marking errors (i.e. marking items that are correct as incorrect, or vice versa) nor clerical errors (i.e. incorrect adding up of scores) could occur. In addition, a random sample of 25% of the papers was to be checked a third time by a designated member of the board.

8.4.9 *COMPUTER ANALYSIS OF RESULTS*

No detailed computer analysis was to be carried out before reporting marks. This was because the examination was not computer marked, and the volume of data could not be processed in the short time available. The only statistics that we were able to make available to the Examination board, in the case of the full administration of sub-tests es01.2, es02, in02, and in03, were the basic descriptive statistics, and the results of the z-value analysis in the case of the parallel tests es01.2 and es02.

After the examination sessions had been concluded, as a part of the ongoing development of the tests, candidates' answers were coded in order to facilitate entry to the spreadsheet software according to the model described by Frude (1993). Copies of all of the codebooks appear in the Appendices. This data was entered onto a PC spreadsheet (Borland 1995) for subsequent analysis.

8.4.10 *STATISTICAL ANALYSIS: CLASSICAL TEST THEORY (CTT)*

Test level analysis of each of the tests involved producing descriptive statistics and a frequency distribution graph and split-half reliability was calculated. Item level analysis, within the bounds of Classical Test Theory (CTT), was carried out to produce Facility values (FV) and Discrimination indices (DI) for each of the items, and distractor analysis data for all of the options. These were computed on the basis of the key responses and the actual responses. Revision of the answers then took place sometimes giving rise to an adjusted markscheme.

Discussion of item performance was based on the following table in which crude verbal labels are given to bands of FV and DI scores in order to facilitate description.

Table 8—8 Terms used to describe item performance (Cohen and Mannion 1989; Heaton 1988)

FV range	*Degree of difficulty*
0.81-1.00	Very easy
0.61-0.80	Easy
0.41-0.60	Average
0.21-0.40	Difficult
0-0.20	Very difficult
DI range	*Value of item*
>0.40	Very good
0.30-0.39	Good
0.20-0.29	Average
(DI)-0.19	Deficient

We established our two FV targets for item performance, the first was for the "easy items" with which we would begin the papers, and the second was for the other items.

Table 8— 9 Target FV levels of item performance established for our sub-tests

Easy	[illegible]0.80
Normal	0.30-0.70

The target DI values for all items were set at the same level, regardless of their FV.

Table 8— 10 Target DI level of item performance established for our sub-tests

DI	[illegible]0.30

8.4.10.1 *Internal consistency*

Given the assumptions required in the use of the principal formulae for the calculation of internal consistency, and mindful of Bachman's exhortation that

> *...in situations where it is not possible to retest individuals, or to administer equivalent forms, and some estimate of internal consistency is the only means of examining the reliability of the test, we must make every attempt to split the test into halves in such a way as to maximize their equivalence and their independence (Bachman 1990:174)*

we opted to calculate split-half reliability with both the SPEARMAN-BROWN and the GUTTMAN split-half formulae. Furthermore, we divided the test items twice, using different criteria, in order to "maximize their equivalence". To do this, we rank-ordered the items by FV score first, and the by DI values. We used these lists to assign items to halves following *the odd-even* principle. Finally, we calculated a set of four coefficients, and the averages of these. This enabled us to target our efforts, but at the same time to cast the net as wide as possible, in line with Bachman's proposal (175). The figure that we finally obtained was, then, the average of the four averages, and it was this data that we later used in further calculations, such as THE SPEARMAN-BROWN PROPHECY FORMULA.

The following table illustrates the full set of calculations that we have just described.

	Split-half coefficient formulae		
Criterion	Spearman-Brown	Guttman	Average
Facility value	x	y	$r_1 = \frac{(x+y)}{2}$
Discrimination value	w	z	$r_2 = \frac{(w+z)}{2}$
Average	$r_3 = \frac{(x+w)}{2}$	$r_4 = \frac{(y+z)}{2}$	
Average overall	$r_0 = \frac{(r_1+r_2+r_3+r_4)}{4}$		

Our target was to obtain a satisfactory level of internal consistency in order to ensure the reliability of our sub-tests and a reasonably low level for the Standard error of measurement.

Table 8– 11 Target level of test performance

Split-half coefficient of reliability	≥ 0.90
Spearman-Brown Prophecy formula	<1

8.4.11 *REPORTING SCORES AND SETTING PASS MARKS*

In this section, the process of fixing the Pass/fail boundary will be described. It can be seen that the political considerations of the process do in fact mediate in favor of this study because the range of candidates admitted to the degree program did not depend solely on the test itself.

The entry test in itself was not the only instrument by which candidates were measured, and it was not necessarily decisive in the entry process. For a number of considerations, which are outside the scope of this study, candidates were not rank ordered in terms of the scores they obtained in the entry test, but were simply divided on a Pass/Fail boundary. In practice, candidates' entry test scores were scaled by a factor derived from their University entrance examination scores and ranked in order of this combined score. This meant that candidates with low scores in the entry test, and high University entrance examination scores might be accepted to the program in preference to others with very high scores in the entry test.

The procedure adopted by the University administration involves equating a "Pass" in the entry test with a score of 5.5. They then calculate the average attained by the individual candidate when adding this to their mark for the University entrance examination. In this manner, no discrimination between candidates is made because of the aptitude test. Having passed the examination, all were considered equally "apt". The weighting of the entry procedure fell in favor of the wide-ranging *selectividad* examination.

Again, there are political and ethical consequences of this procedure which, although not of direct concern to our research, do influence its results. In the context of a degree program for which there is a limited number of places — *numerus clausus* — as is the case in Translation and Interpreting, if the pass/fail boundary for the entry test is set so as to pass only the same number of candidates as there are places available, then the *selectividad* score does not enter the equation. If, however, the pass/fail boundary is set so as to pass a number of candidates greater than the number of places available, then the *selectividad* score is decisive, over and above that of the entry test. The greater the number of candidates considered to have passed the more important their scores in the *selectividad* examination become.

The decision on this aspect of the entry test is left in the hands of the Examination board, but clearly, it has far-reaching consequences. Not the least of these, from the perspective of our research, is that fact that the students entering the program do not represent those with the highest levels of aptitude for translation. This approach means that the function of the entry test is devalued, and that the prospective students, who have paid a substantial fee to take the test, are left to question its academic purpose.

8.5 Sampling

In this section we will look at the general question of sampling in research design, and at the specific problems which arise in aptitude testing in "naturally occurring settings" (Larson-

Freeman and Long 1991:22) such as ours. We will describe the manner in which we propose to handle these problems.

8.5.1 *POPULATION*

Our entry test is designed for the population of "potential candidates for places on the first degree program in Translation and Interpreting". At the design stage of our study, we believe that the cohorts we plan to use for trialling and for testing, are more or less representative of this population. The initial sub-tests were to be trialled on students who had already gained entry to the Diploma course in Translation and Interpreting, and the full tests were to be administered to genuine candidates made up all of those applying to start the course in the academic years 1993-94 and 1994-95. The starting points for our research are, therefore, largely representative, and we believe they provide a good basis for generalizing about the target population.

8.5.2 *SAMPLE SIZE*

The longitudinal aspect of our study is subject to problems arising from the administration of instruments to subsequent cohorts of individuals derived from the initial groups over a period of time up to two years, and on the composition of these cohorts.

It is a paradox that, in research terms, one of the difficulties in measuring aptitude is that real world investigation entails accepting a basic design flaw. Aptitude testing in modern languages teaching was and is used as a means of selecting learners. Its purpose principally is "to keep prospective failures out of classes" (Spolsky 1995:117) and thus to minimize what Cheydleur (1932, quoted in Spolsky 1995:117) called the "mortality of modern languages students". This very "mortality", however, afflicts the samples under study, as part of the initial group is rejected at the first hurdle, and therefore is not available for subsequent analysis.

In addition to this basic design flaw, sample erosion cannot help but influence results. With a longitudinal study of this nature, covering a period of two academic years, the additional "mortality" of the cohorts is also likely to be high because of the very nature of Translation studies programs. Almost all of the students take up Erasmus/Socrates grants at some stage in their undergraduate studies and may spend a full academic year abroad. Many also spend summers abroad specifically in order to improve their language skills. Others also supplement the teaching they receive within the university by attending extra language training courses.

Sample erosion over the two years of our data collection period is, therefore, an obvious threat to this study. Hicks (1990) questions the need for samples that are greater than 25, and suggests that the choice of statistical analysis can compensate for sample size difficulties. She proposes the following guidelines:

Table 8—12 Minimum sample size (Hicks 1990)

N° of subjects	
Minimum	12
Acceptable	20-25
Optimum	25

In a similar vein Kendall (1975) offers worked examples based on samples of 12, and the calculation of levels of significance for results are based on tables which use 30 as the largest group size (Kendall 1975; Siegel and Castellan 1998; Silver 1997). These arguments notwithstanding, in this study we will work with the largest possible samples in order to ensure that erosion does not reduce the cohorts available to us for the final measures to < 30, as suggested by Rowntree (1981).

8.5.3 *SAMPLING ERROR*

To this end, within the period of longitudinal testing, a process of conscious sampling in order to maintain the appropriate groups was rejected. In addition to our wish to avoid sample erosion, two other factors influenced our decision. Firstly, conscious sampling was thought likely to produce a "control effect" (Cohen and Mannion 1989:73) whereby some candidates might prepare especially for the tests in an attempt to perform as well as possible; and others might simply stay away. In either case, this would produce a distortion in the results with respect to other candidates who are in a better or less well-prepared state.

Secondly, the management of this process would have required a certain amount of cooperation from members of other departments who could not always be counted on to share the same interest in the successful administration of the tests. These factors led to the acceptance of a "convenience sample" to be measured with respect to the whole group to try to establish its representativeness. The smaller the resulting sampling error mean, the more representative the figures can be said to be.

8.5.4 *SOURCES OF ERROR*

One predictable source of information loss, though not actual error, is the nature of the data to be collected. While the data over which we have direct control is, by design, interval level data — i.e. numerical scores out of 20 — the data that we will collect from the General Translation course may not necessarily be of the same type. Correlations between ordinal data — i.e. rank ordered grades: *suspenso*, *aprobado*, *notable* or *sobresaliente*, — which we expected to collect, and interval data cannot be calculated using the Pearson coefficient. We therefore prepared to use KENDALL'S TAU$_{B}$ (T_b) coefficient, which works on the principle of correlating the rank order of subjects and compensating for the tied cases. This involves calculating how many changes would need to be made in a given list for the order in which the individual candidates appear to be the same as that in another list. T_b offers a precise calculation when dealing with a large number of tied cases, the most predictable outcome of applying a limited number of ordinal measures to a large sample.

8.6 Validation and test analysis

We were concerned that the experimental nature of our entry test might actually produce results that would prejudice candidates and/or negatively influence the composition of the actual first intake of students. In order to check on this we planned to administer the Oxford Placement Test (OPT) in order to compare those students who took the entry test, with those who had entered the Diploma course in previous years. Earlier in our study, we have described the use of this test as part of a process of measuring the actual levels of English language competence among Diploma course students, and now we will describe the test in more detail.

8.6.1 *OXFORD PLACEMENT TEST*

Published by Oxford University Press (Allan 1990), the OPT aims to test and discriminate between all levels of competence from elementary to post-Cambridge Proficiency English.

The test is in three parts: one Listening sub-test, and two Grammar sub-tests. Relative weighting gives 50% of the raw score (max = 200) to the Listening sub-test, and 25% each to the Grammar components.

The Listening sub-test assesses both reading and listening skills, and indicates candidate ability to manage both the sound and the writing systems of English. The Grammar component tests knowledge and application of grammatical structures.

The OPT is commercially available from Oxford University Press and is reported by its author (Allan 1990) to produce "very high" correlations with scores on other test batteries. Allan gives the figure 0.89+, although he does not indicate which test batteries he has used. Validation of the OPT involved five years of research around the world. The test instructions include a table of equivalence, from which candidate OPT scores can be used to approximate their level of ability for placement and for prediction of exam results against eight sets of examinations run by British-based examining boards. We have already presented an adapted version of this table on page 31.

We chose the OPT to validate (English) proficiency because of the wide range of levels covered, the table of equivalence with exams familiar in Spain, the ease of administration in terms of time and correction, and the academic rigor with which the test has been prepared.

8.6.1.1 *Analysis of the OPT Listening/Reading sub-test*

Source material is derived from native speaker usage and was trialled on native speakers. A detailed process of item analysis produced Facility values and Discrimination indices, which were used to refine the test. Item and inter-test reliability were also tested as was concurrent validity with other batteries of tests.

The listening component is based on semi-authentic, single sentences chosen for their sound similarities only. The items are not connected in any way.

There are 100 items, all of which are of the dichotomous task type, based on a simple choice. Candidates are required to identify the written form that represents what they have heard.

8.7 Conclusions

As a result of our study of the research design process, we have established a series of criteria that we will use in order to evaluate the performance of our test. These criteria, presented in Table 18—13 are shown with the experimental or statistical means by which we intend to evaluate them, and the target standards we believe our test instruments should meet.

The criteria are divided into four basic sections, following the order in which they are defined earlier in Chapter 7. These are validity, reliability, and practicality, followed by the two elements of CTT item analysis: Facility value and Discrimination.

Table 8 – 13 Summary of test implementation criteria, means of testing, and targets

Sub-test performance criterion	*Means of testing*	*Target parameters*
Concurrent validity	Correlation coefficients between LA and LB sub-tests and OPT, University entrance test, and Secondary school average	≥0.65 at 5% significance
Construct validity	Correlation coefficients between LB sub-tests and OPT	≥0.65 at 5% significance
Content validity	Expert opinions offered by members of the examination board	Reference to the test specification
Criteria validity	Correlation coefficients between LB sub-tests and OPT	≥0.65 at 5% significance
Face validity	Informal feedback from (successful) candidates	
Response validity (Carroll's condition 1)	Standard error of measurement	<1
Predictive validity (Carroll's condition 6)	Correlation coefficients between LA and LB sub-tests and General translation A-B and B-A	≥0.65 at 5% significance
Reliability	Average of split-half coefficients	≥0.9
Practicality	Correlation coefficients between LA and LB sub-tests and OPT, University entrance test, and Secondary school average	≥0.65 at 5% significance
Facility values	0.3-0.7	
Discrimination indices	≥0.3	

9 INSTRUMENT DESI GN

9 INSTRUMENT DESIGN

9.1 MATCHING EXERCISES
9.2 DICHOTOMOUS ITEMS
9.3 MULTIPLE CHOICE QUESTIONS
9.3.1.1 The random factor in marking objective tests
9.4 ERROR IDENTIFICATION AND ERROR CORRECTION
9.5 SUMMARY WRITING
9.5.1 *Summary writing exercises used by centers in Spain*
9.5.1.1 University of Alicante
9.5.1.2 Autonomous University of Barcelona
9.6 UNIVERSIDAD PONTIFICIA DE COMILLAS
9.7 CONCLUSIONS

IN THIS CHAPTER we intend

- to evaluate test task types for inclusion in our test
- to describe and discuss examples of entry tests used by other centers in Spain

9.1 Matching exercise s

This test task involves candidates connecting items in two or more lists, or from a list with reference to a text (Table 9—1). The basis of the connection is very often based on the skill of *"deducing the meaning and use of unfamiliar lexical items"* (Munby 1978:126-131), although it will involve the use of other skills in order to identify the parts of the text where the response lies. In fact, skimming and scanning skills are essential for candidates to resolve these items, and in the simplest form, simple or complex scanning are the only skills that are tested.

Table 9— 1 An example of a matching exercise (Robinson 1999)

In the subtitle there is a word that can be defined as "someone who has been treated badly by those in power". The word is *downtrodden*

Read these definitions. Then read the text to find the word or phrase that matches each, and write your answers in the spaces provided.

1 to suffer something unpleasant until it ends
2 Saved
3 moved on to something better
4 a very successful production
5 Executives
6 first performance
7 gained fame
8 made as perfect as possible
9 the high point

The most important design aspect of this item type is the ratio of stem prompts to alternative answers. A simple form of the task, such as the linking of words with their corresponding definitions, can have a negative effect on candidates if the number of prompts and the number of alternatives is the same. This negative effect can take one of two forms: either the final item must be correct by default (Alderson et al 1995:51-52), because there are no alternatives left; or the candidate who makes one mistake, loses two points, because one mistake entails making a second.

The version above is an improved form of the task, as the alternatives can appear anywhere in the text, giving ample scope, and penalizing no one. An appropriate half measure is found by using two lists, but including one alternative more than there are prompts. This ensures that no candidate is penalized by more than one point for making one mistake. The inclusion of more than one extra alternative can also be considered, but an excessive number of alternatives can lead to confusion on the part of the candidates.

9.2 Dichotomous item s

The term "dichotomous items" is used to refer to those task types in which candidates have a simple choice between two alternative responses. These may be "True" or "False", "Yes" or "No", or in more sophisticated items they may be labels of one kind or another "Bob" or "Elaine", for example, when referring to two characters mentioned in the text. The principle difficulty about using these items is that there is a random factor about response, which means that any candidate could score half marks, without actually reading the items. This renders the task type relatively inefficient unless large numbers of such items are used. In the Netherlands, the national testing agency, CITO, uses items of this type for its FL examinations, equivalent to the English "O"- and "A"-levels. Their listening tests involve lengthy passages, of up to 45 minutes in duration, heard once only, and accompanied by 50 or more items in which candidates only have to choose between two alternatives. This test mechanism, the same one as that used in the *Oxford Placement Test* (Allan 1990), is considered highly effective.

Table 9— 2 Some (true/false/questionable) items (Robinson 1996)

The following statements are based on the content of the text. Some/Three of them are TRUE, some/three are FALSE, and some/four are QUESTIONABLE, that is they are not supported by information which appears in the text. Two examples have been given for you.

1 The arrival of a democratically elected government in South Africa has given a boost to black theater ______

2 Previously, black theater had revolved around the theme of apartheid. ______

3 Ngema has stated that he will continue to act as the conscience of the A.N.C.-led South African government. ______

In order to introduce a third element to the basic true/false model, and in this way reduce the random factor, various authors propose the true/false/questionable or don't know task (Heaton 1988; Carroll Brendan J. 1993).

The item type gives candidates the choice between that which is demonstrably true, from the information given in the text; that which is demonstrably false; and that which cannot be proven one to be either true or false because no information is given about it.

9.3 Multiple choice questions

Multiple choice questions, whether with three, four, or five options, are often thought to be the most adequate means of testing reading skills because candidates barely have to use any skills that are not identifiably based on reading. The written response to these items is usually a line, cross, circle, or ✓; all else depends on the reading skills activated by the candidate in accessing the stem, options, and text to which these refer.

Table 9—3 Some examples of four option multiple choice items (Robinson 1996)

1 When people communicate by using a sign language they present their message
A in linear fashion. ____
B by fingerspelling. ____
C through facial expression. ____
D using spatial relations. ____

2 Sign languages around the world
A have evolved in parallel. ____
B are, in part, mutually comprehensible. ____
C reflect conventional languages. ____
D differ only in grammar and syntax. ____

3 Sign languages are similar to conventional languages in that they
A Can be transcribed in written form. ____
B Have different regional dialects. ____
C Use the same basic grammar. ____
D Are of about the same age. ____

However, one of the questions that arises in the use of MCQ items for reading comprehension is the matter of whether candidates are required to respond from their reading of the text, or whether they can respond only from their reading of the item. Clearly, some skills adapt to the MCQ task type better than others do. Alderson et al (1995:45) suggests that for inferencing, MCQ items, while extremely difficult to write, are perhaps the best choice as they can potentially "guide" the reader in their interpretation of the text.

A recent study of the limitations of MCQ items concludes that the construct validity of the format is open to question. In a study of Chinese learners' test-taking processes in a listening comprehension test, Wu Yi'an (1998) used retrospection protocols to go into the mental processes lying behind candidate performance. The results suggest that the format favored more advanced learners and provided added difficulty for the less able.

This conclusion was expanded in four parts:

★ The item stems and the options helped advanced learners in that these focussed their listening; less able learners were not helped.

★ Construct validity was "threatened" as the level of difficulty in the options was much more complex and converted the items themselves into a more demanding test of vocabulary and reading comprehension.

★ "Uninformed guessing" took place. This was guesswork based on candidates' non-linguistic knowledge, which they activated in order to cope with their partial linguistic processing of the text. In less able candidates it gave rise to "the lure of options", which lead them away from processing the linguistic information of the test input.

★ The right answers could be given for the wrong reasons.

From the point of view of our reading comprehension test, these negative consequences might not all apply. In a reading test, we assume that all input is part of the testing process, and consequently this kind of interference, which might threaten validity, does not apply. Reading comprehension is our objective, and vocabulary testing is an important aspect of this. However, construct validity would begin to be questioned if candidates were required to employ different skills than those targeted by the items, or if the level of difficulty of the items, and/or rubric, were greater than that of the text.

9.3.1.1 *The random factor in marking objective tests*

We have earlier discussed the question of the random factor and how it affects dichotomous items and others, such as these four-option MCQs. One of the procedures used to compensate for this factor is described by Alcaraz Varó and Ramón y Denia

(1980:69). The authors describe different formulae in which the candidates are scored by subtracting the number of incorrect items from the number of correct ones. In each case, the proportion of the incorrect items subtracted is relative to the number of options in the item. For example, if a candidate scores 6 out of 10, for five-option MCQs, the final score awarded would be calculated by subtracting one quarter of the incorrect items, i.e. 1, from that score, giving 5 out of 10.

Table 9– 4 A formula to compensate for the random factor in five-option MCQ items

Total n° of MCQ items	=	10
N° of options	=	5
Items correct (c)	=	6
Items incorrect (i)	=	4
i / 4	=	1
Candidate raw score	=	5

In this way, guessing is penalized. However, as a procedure it is one that must explained to candidates in advance of the test administration in order to ensure that no one is disadvantaged by the use of this scoring mechanism.

9.4 Error identificatio n and error correction

These test tasks are described by Heaton (1988 & 1990) and by Weir (1988), both of whom classify them as tasks used in testing writing skills. Heaton uses the term "error-recognition", and Weir talks about "editing skills", when they describe a range of tasks, all of which are variations on the same basic activity, ranging from multiple choice activities to more authentic text editing.

Heaton describes three tasks that involve identifying the location of an error, and in one case correcting the error. In each of the first two types a sentence containing one error is used as the stem. Four words or phrases from the sentence are underlined and labeled in multiple choice fashion: A, B, C, or D. Candidates then identify the phrase containing the error by choosing the appropriate letter.

It Was a terrible accident at an air-show
A B C
held In West Germany yesterday.
D

This is highly specific, and an alternative that Heaton appears to prefer is described (1990:98)

It was / a terrible accident / at an air-show
A B C
/ Yesterday.
D

Here the candidates again have to choose the part of the sentence that contains the error, but in a less directed manner. Heaton suggests that this method is useful for testing problems of omission (1990:98) or spelling (1988:152).

The third alternative involves presenting sentences, such as the example, which contain one error each, and asking candidates to write out a correct version of the sentence. The disadvantage of this type lies, he affirms, in the marking. Many questions arise, such as whether or not to penalize the candidates' versions that contain errors other than the target error.

Weir (1988:59-60) discusses these item types and mentions the advantages they bring as they are more productive, and that they meet construct validity criteria. However, he too stresses the problems that lie in marking items of this kind.

This task type is used by examining boards, and Weir (1988:146, 206-207) presents examples taken from their papers. The *Test in English for Educational Purposes (TEEP)* sample gives a reasonably authentic version of this item type in which a passage is presented, and candidates are asked to identify errors by underlining them, and then to write a correction beneath the error in each case. The number of errors in the passage is not specified, but the types of error are. These cover grammar, spelling, and punctuation.

In the JMB *University entrance Test in English for Speakers of Other Languages (JMBTesol)* which Weir also reproduces here, we find tasks in "Editing Skills", again based on a complete passage, but in which candidates are informed that one word has been left out of each line. They are asked to identify where the word is missing by marking the position with a line, and writing the missing word in the space provided in a parallel column.

The same paper contains another exercise which is an improvement on the *TEEP* in that it specifies there is only one mistake per line, and that there are no spelling mistakes.

9.5 Summary writing

Newmark (1988) makes the point that summary writing in itself is an excellent learning task in the training of translators. This view is clearly that shared by those teachers of translation who adopt the summary exercise in one form or another as a means of testing for aptitude. There are many variations on the basic task, but essentially it is employed along the lines we have described earlier (Chapter 3). The task type has high face validity, and can focus on the integrated use of a range of reading and writing skills. The principle disadvantage, as we have shown, lies in the difficulty of ensuring reliable and consistent marking. Weir underlines the importance of detailed preparation of the markscheme (1990:63), and the inevitability of a degree of subjectivity. Notwithstanding these drawbacks, he believes that with an adequate investment of time, preparation, and rater training, it is possible to achieve acceptable levels of reliability.

9.5.1 *SUMMARY WRITING EXER CISES USED BY CENTERS IN SPAIN*

In January 1994 in preparation for the *1er conferencia inter-centros* held at the University of Granada, participants were asked to respond to a number of questions relating to the entrance test. Of those attending, both the University of Alicante, and the Autonomous University of Barcelona (*UAB*) reported they had adopted a summary writing test of one kind or another as part or all of their entry procedures. We will now describe the procedures they used and discuss some of the points arising.

9.5.1.1 *University of Alicante*

In Alicante testing involved two sub-tests. Firstly, candidates were required to write summaries in one or other of the official languages of the autonomous community: Spanish or Valencian. Their summaries were based on listening to a talk, which was to last a maximum of 30 minutes. Additionally, they were tested on their ability to translate, without the aid of a dictionary, a general text of 500 words maximum from their chosen Language B into either Spanish or Valencian. We find it difficult to see how this second sub-test could be considered legal given that the interpretation of the official requirements accepted by other centers precluded the use of actual translation exercises on the grounds that these tested skills candidates would be taught on entry to the university.

9.5.1.2 *Autonomous University of Barcelona*

The Autonomous University of Barcelona (*UAB*) used a variation on the LB-oriented, integrative model described above, since the introduction of the degree program. At the time of the 1994 conference, it was reported that the results of these tests had been monitored *ad hoc* and unspecified changes made to increase "efficiency". Since then more detailed, but unpublished, research has been carried out (Fox 1997).

In the summer of 1995, the *UAB* introduced an LA (Catalan or Spanish) sub-test into the entrance examination for the first time. The general conclusion of teachers in the light of translation examinations at the end of the first semester 1995-96 was that this had brought about an improvement in the "quality" of the intake (Beeby Lonsdale 1996).

This examination was in two parts: LA (Spanish or Catalan), and LB (English French, or German). The LB (English) test was based on an 800-word, amended transcript of a television documentary dealing with a topic of current interest. Amendments were made to the text in order to compensate for the lack of visual images. This was read aloud by an LB native speaker, with as natural an intonation as possible. It was read twice, and candidates were then allowed an hour in which to write a 300-word summary. This was corrected according to pre-established criteria.

These criteria centered on accuracy of content, with positive points up to a maximum of 10 for including the predicted "main points" of the text; and accuracy of grammar/syntax, with negative half points for mistakes, again up to a maximum of 10. The nature of errors was predicted in a list provided for members of the Examining board.

The marking of this summary writing exercise produced a number of problems of the type Weir had warned about (1990). Raters found that they needed to adjust the correction criteria to deal with two kinds of candidate who did not fit the predicted profile. Firstly, there were those candidates who wrote short, grammatically correct texts including enough of the main points to pass, but who did not demonstrate they were "discourse competent". Secondly, came those who were judged to be "discourse competent", but who failed due to the number of linguistic errors they made. A definition of "discourse competence" was not given, nor has any statistical analysis of these tests been published. However, the question of intra and inter-rater reliability are currently under discussion (Fox 1997), and it is evident that implicit in the debate is the establishment prior to testing of an agreed minimum standard for entry. We interpret these findings as further evidence of the difficulty involved in the use of subjective exercises of this type, and of the inherent problem arising when tests are prepared on an uncertain theoretical basis.

9.6 Universidad Pontificia de Comillas

We will now comment briefly on the entry test used at this private university. Our motive for doing so is the fact that the entry test they employ contrasts totally with those of other centers, and involves a series of six sub-tests, five of which are objectively marked. These sub-tests were:

- 100 item multiple choice test of LB grammar
- 30 item multiple choice test of LA
- 30 item vocabulary test (LA)
- Intelligence test (LA)
- General knowledge test based on the secondary school curriculum
- Short translation from LB into LA.

This last exercise was the only part of the testing process not machine marked.

All of these sub-tests had been prepared over a period of years involving a good deal of research and pre-testing. The University was satisfied that the correlations of scores produced on these tests were acceptable. None of this information had been published at that time. (Waddington 1994. Personal communication) In addition, candidates successful in the formal tests attended a series of ten-minute oral interviews conducted in LB, LC, and LA (Spanish) respectively. A fourth interview, in Spanish was of an informative nature, and gave them the opportunity to question interviewers about the degree program.

9.7 Conclusions

In this chapter, we have looked at design aspects of the instrument that we intend to use for our research. The survey has by no means covered all of the task types that are described in the literature, but it has concentrated on those that we consider the more appropriate. Our criterion for this has been the practicality of using and marking the tasks in our bid to attain maximum objectivity. Thus, the summary writing task that we describe is not used in any of our sub-tests due to the difficulties of achieving objective, reliable marking, as we have described in Chapter 3.

Our first draft test specification includes a range of these test types, and this was to be revised after administering the trial version to a cohort of Diploma course students. Further modifications were planned after analysis of the first batch of test results, from Cohort A.

Table 9—1 shows the way in which the design process drew together the test specification list of skills and the selection of test tasks in order to write the test paper.

Table 9—1 An example of the design process (sub-test es01.2)

Item	*Task type*	*Skills (Munby 1978:126-131)*
#1	Matching exercise	Scanning to locate specifically required information on a whole topic (Skill 46.5)
#2		Idem
#3		Idem
#4		Idem
#5	True/false/ no evidence	Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)
#6		Idem
#7		Understanding relations between parts of text through the grammatical cohesion device of reference (Skill 32.1)
#8		Understanding explicitly stated information (Skill 20)
#9		Idem
#10		Idem
#11	5-option MCQ	Understanding relations within the sentence (Skill 28)
#12		Idem
#13		Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)
#14		Understanding explicitly stated information (Skill 20)
#15		Understanding relations between parts of a text through grammatical cohesion devices of reference (anaphoric and cataphoric) (Skill 33.1)
#16		Understanding explicitly stated information (Skill 20)
#17		Deducing the meaning and use of unfamiliar lexical items, through understanding word formation, and context (Skills 19.1 & 19.2)
#18		Idem
#19		Idem
#20		Idem

Part III

Results

10 TEST DEVELOPMEN T AND REFINEMENT

10 TEST DEVELOPMENT AND REFINEMENT

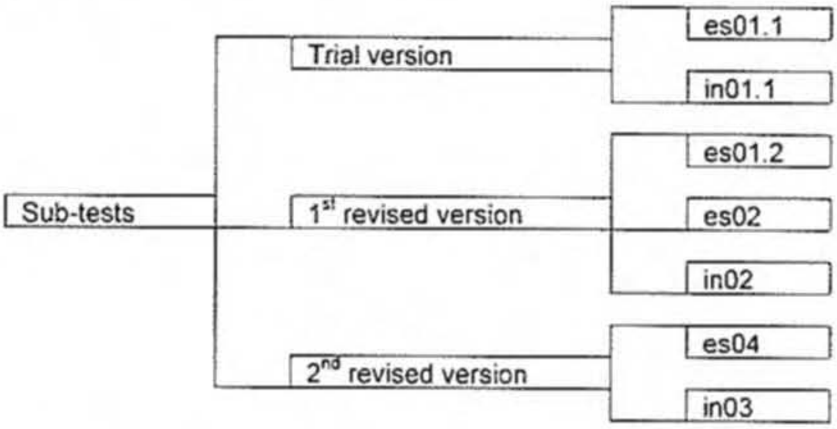
10.1 THE TEST DEVELOPMENT AND REFINEMENT PROCESS

10.2 A NOTE ABOUT PRESENTATION

10.1 The test development and refinement process

This chapter introduces a series of seven, of which each is dedicated to the analysis of one of the sub-tests we have prepared. The development of the tests took place in three stages as shown in Table 10—1. The initial trial phase involved two tests, and this was followed by a second and a third stage each of which incorporated modifications in the light of previous sub-test performance.

Table 10—1 Stages in the sub-test development process

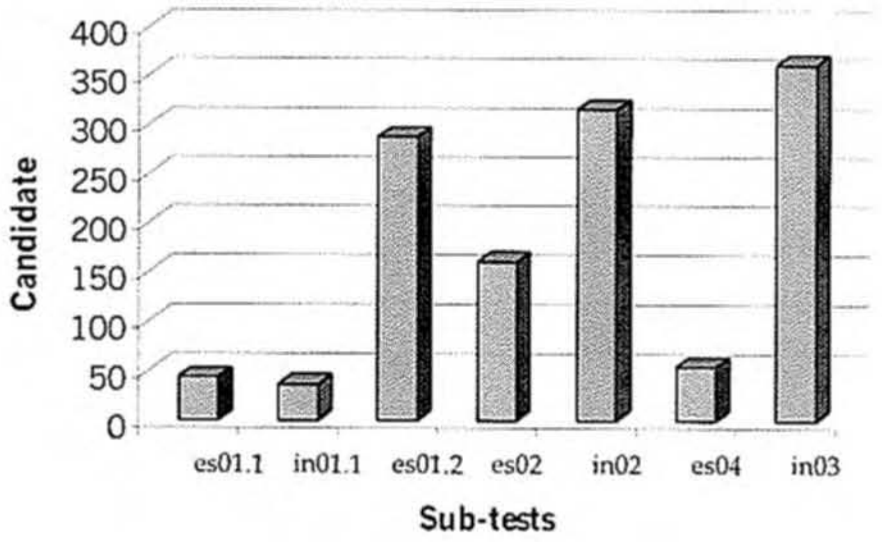


The papers fall into three groups, which we discuss in the order in which they appear in the table, that is the chronological order of administration. First come the trial sub-tests es01.1 and in01.1, which were taken by a sample of Diploma course students. Secondly, we describe the first revised versions of the sub-tests administered to candidates in July 1993 (sub-tests es01.2, es02, and in02). Finally, we look at a second revision of the design, which was used in July 1994 (sub-tests es04 and in03).

In line with our test design objective, we have managed to ensure a statistically sound minimum of >30 candidates for each sub-test. However, the validity of the statistical data we present must always be considered in the light of the actual number of candidates involved in each instance. The sometimes large differences in cohort size between one test and another are illustrated in Figure 10—1.

In the course of these chapters we describe the results obtained paper by paper following the general outline shown in the table below. In each case we begin with comments about the performance of the sub-test. We follow this with a brief introduction to the sub-test itself, in which we describe the context in which it was administered. We then print a word-

Figure 10—1 A comparison of the numbers of candidates in the cohorts of each of the sub-tests analyzed



processed version of the text used in the paper prior to beginning on our analysis of item performance.

Table 10—2 Summary of contents: these are the elements that typically appear in each of the following chapters

General comments	
Text	
	Exam text
Task types	
	Facility values (FVs)
	Discrimination indices (DIs)
	Discussion
	Distracter evaluation
Descriptive statistics	
Reliability	
Conclusions	

In order to contextualize the results of our analysis we describe the task type and target reading skills, and reproduce the items in each section when relevant. We present the FACILITY VALUE (FV) and DISCRIMINATION INDEX (DI) scores for each of the items of the task type in question, using graphics and tables in the presentation of data to highlight key information, but at the expense of detail.

In our discussions of item performance we indicate which items performed within our target parameters, evaluate the performance of distractors, and pay particular attention to those that were discarded after analysis. We provide calculations of reliability using the KUDER-RICHARDSON FORMULA 20 (KR 20), or the SPEARMAN-BROWN and GUTTMAN split-half formulae, calculated according to both the FV and DI rank orders of items. We also give the results of calculating THE SPEARMAN-BROWN PROPHECY FORMULA and the STANDARD MEASUREMENT OF ERROR (S_e).

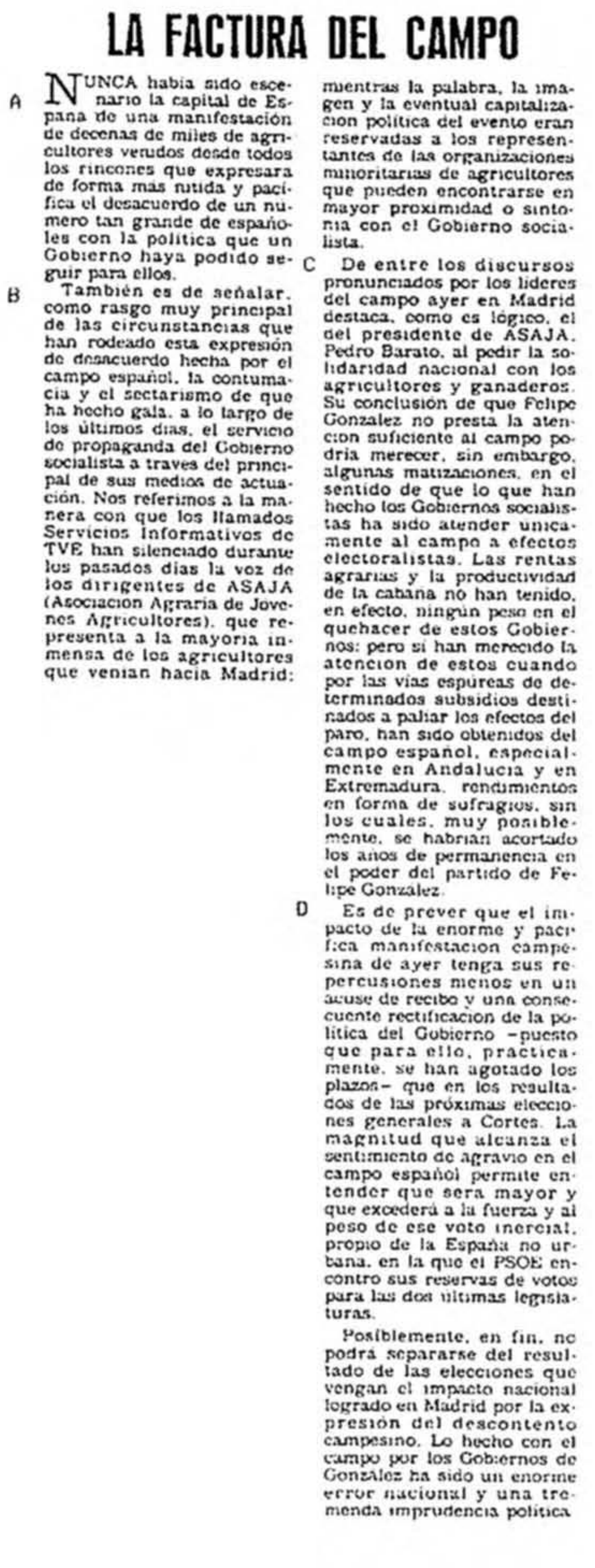
We present histograms of the frequency distribution of scores achieved by candidates, scaled when necessary to totals out of a possible 20 in order to facilitate comparison between sub-tests. And we also offer a summary of some of the descriptive statistics for each paper: mean, median, mode, standard error, standard deviation, skewness, variance and confidence level (calculated at 95%).

Our conclusions about each test deal with the specific aspects of overall test performance, and draw particular attention to those results that provided material for the development of the sub-tests described and/or for the preparation of further papers.

10.2 A note about presentation

In this part of our work we were faced with the question of how best to present the results and the sub-test papers in order to integrate both into the text without making it disjointed and difficult to read. Having consulted the literature on presenting instructional text (Harley 1994; Lannon 1994) we have chosen to present only one of the sub-tests in full — sub-test es01.2, which begins on the next page — in order that it may serve as an example of the manner in which all sub-tests were produced. For each of the others, we have reserved the full version for the Appendices, where those who have a specific interest in one or other test can consult them. Here we will reproduce in word-processed format a full version of the text only. The specific items will, however be reproduced — also in word-processed format — as the results derived from them are described.

We realize that this decision is a compromise, but feel that it will provide readers with the least cumbersome approach to the sub-test results.



LA FACTURA DEL CAMPO

A NUNCA habia sido escenario la capital de Espana de una manifestación de decenas de miles de agricultores venidos desde todos los rincones que expresara de forma más nitida y pacífica el desacuerdo de un número tan grande de españoles con la política que un Gobierno haya podido seguir para ellos.

B También es de señalar, como rasgo muy principal de las circunstancias que han rodeado esta expresión de desacuerdo hecha por el campo español, la contumacia y el sectarismo de que ha hecho gala, a lo largo de los últimos días, el servicio de propaganda del Gobierno socialista a traves del principal de sus medios de actuación. Nos referimos a la manera con que los llamados Servicios Informativos de TVE han silenciado durante los pasados dias la voz de los dirigentes de ASAJA (Asociacion Agraria de Jovenes Agricultores), que representa a la mayoria inmensa de los agricultores que venian hacia Madrid; mientras la palabra, la imagen y la eventual capitalización politica del evento eran reservadas a los representantes de las organizaciones minoritarias de agricultores que pueden encontrarse en mayor proximidad o sintonia con el Gobierno socialista.

C De entre los discursos pronunciados por los lideres del campo ayer en Madrid destaca, como es lógico, el del presidente de ASAJA, Pedro Barato, al pedir la solidaridad nacional con los agricultores y ganaderos. Su conclusión de que Felipe Gonzalez no presta la atencion suficiente al campo podria merecer, sin embargo, algunas matizaciones, en el sentido de que lo que han hecho los Gobiernos socialistas ha sido atender unicamente al campo a efectos electoralistas. Las rentas agrarias y la productividad de la cabaña no han tenido, en efecto, ningún peso en el quehacer de estos Gobiernos; pero si han merecido la atencion de estos cuando por las vías espúreas de determinados subsidios destinados a paliar los efectos del paro, han sido obtenidos del campo español, especialmente en Andalucia y en Extremadura, rendimientos en forma de sufragios, sin los cuales, muy posiblemente, se habrian acortado los años de permanencia en el poder del partido de Felipe Gonzalez.

D Es de prever que el impacto de la enorme y pacífica manifestacion campesina de ayer tenga sus repercusiones menos en un acuse de recibo y una consecuente rectificacion de la política del Gobierno –puesto que para ello, practicamente, se han agotado los plazos– que en los resultados de las próximas elecciones generales a Cortes. La magnitud que alcanza el sentimiento de agravio en el campo español permite entender que sera mayor y que excederá a la fuerza y al peso de ese voto inercial, propio de la España no urbana, en la que el PSOE encontro sus reservas de votos para las dos ultimas legislaturas.

Posiblemente, en fin, no podrá separarse del resultado de las elecciones que vengan el impacto nacional logrado en Madrid por la expresión del descontento campesino. Lo hecho con el campo por los Gobiernos de González ha sido un enorme error nacional y una tremenda imprudencia política

"La factura del campo"

Lee el texto y señala a cuál de los párrafos podría corresponder cada una de las siguientes frases como título. Subraya la letra en la hoja de respuestas.

1. "El acontecimiento".
2. "La actitud del gobierno hacia el campo".
3. "El papel del Telediario".
4. "Las próximas elecciones".

V (Verdadero), F (Falso) o ? (el texto no informa sobre este punto), y subraya la letra o el símbolo correspondiente en la hoja de respuestas.

5. Es muy significativo que esta manifestación haya tenido lugar en Madrid.
6. Pedro Barato opina que los socialistas perderán las próximas elecciones.
7. Los servicios informativos de TVE siempre se muestran parciales.
8. Si no fuera por el voto del campo el PSOE nunca habría llegado a gobernar.
9. En opinión del autor de este artículo el campo seguirá apoyando al Gobierno a pesar de lo hecho.
10. Las consecuencias más importantes de la manifestación se verán especialmente en Andalucía y Extremadura.

Elige la opción más apropiada y subraya la letra en la hoja de respuestas.

11. En opinión del autor de este texto, la conclusión de Pedro Barato es ...
    A correcta.
    B incompleta.
    C electoralista.
    D suficiente.
    E superflua.

12. Como resultado de la presentación de esta noticia por parte de TVE __________ se ha(n) visto perjudicado(s).
    A ... ASAJA ...
    B ... organizaciones minoritarias de agricultores ...
    C ... ningún colectivo ...
    D ... todos ...
    E ... el PSOE ...

13. Una rectificación por parte del Gobierno ...
    A influiría en las elecciones.
    B cambiaría la política de TVE.
    C sería muy bien recibido.
    D tendría efectos retroactivos.
    E llegaría demasiado tarde.

14. ¿Qué es lo que el Gobierno ha obtenido?
    A Subsidios.
    B Votos.
    C Beneficios.
    D Rentas agrarias.
    E Apoyos.

15. ¿Qué será mayor?
A La mayoría del nuevo gobierno.
B El sentimiento en contra del PSOE.
C La inercia del campo.
D La próxima manifestación.
E La respuesta urbana.

16. Ninguna manifestación anterior ha ...
A unido las organizaciones del campo.
B obligado al Gobierno a reaccionar.
C recibido tanto apoyo popular.
D destacado la proximidad de Gobierno y agricultores.
E demostrado tan claramente la actitud en contra del Gobierno.

17. La frase "la contumacia y el sectarismo de que ha hecho gala" en el párrafo B, quieren decir que TVE se ha mostrado ...
A neutral.
B irresponsable.
C partidaria del gobierno.
D inconsistente.
E influída por un sector.

18. La frase "la productividad de la cabaña" en el párrafo C refiere a ...

A los impuestos pagados por los agricultores.
B los ingresos que reciben al vender su ganado.
C el volumen de carne producido en España.
D los beneficios del sector ganadero.
E la cuantía de cereales de la cosecha anual.

19. La frase "las vías espúreas" en el párrafo C, son ...

A rutas peligrosas.
B métodos ilegítimos.
C la red viaria.
D un sistema burocrático.
E un planteamiento político.

20. La frase "en un acuse de recibo" en el párrafo D, significa ...

A un formulario oficial.
B una respuesta por escrito.
C una reacción protocolaria.
D un reconocimiento formal.
E una acusación jurídica.

11 SUB-TEST ES01.1

11 SUB-TEST ES01.1

11.1 TEXT
11.2 TASK TYPES
11.2.1 *Matching exercise*
11.2.1.1 Facility values
11.2.1.2 Discrimination indices
11.2.1.3 Discussion
11.2.2 *True/False/Don't know*
11.2.2.1 Facility values
11.2.2.2 Discrimination indices
11.2.2.3 Distractor evaluation
11.2.2.4 Discussion
Item #8
11.2.3 *Sentence completion items*
11.2.4 *Five-option multiple choice questions*
11.3 CONCLUSIONS

THIS WAS the first test paper that we wrote and we were keenly aware of the tensions existing between our lack of experience in writing papers of this type, and our need to produce a successful instrument to be used in a real-world context. We hoped that these would cancel each other out and ensure a professionally written paper.

The test tasks were based on the first draft specification, which appears in an earlier chapter, and included a variety of task types and target skills. Specifically, these were chosen from the range of activities taught and tested in the Diploma course *Traducción I (Inglés)* during academic years 1991-92 and 1992-93. We modelled the task types on materials drawn from a variety of sources in English language teaching and language testing. Among others, these included Moore et al 1979; Barr et al 1981; Grellet 1981; Rudska et al 1981; Carroll, Brendan J. and Hall 1985; Heaton 1988; and Carroll, Brendan J. 1992.

In Spring 1993, 44 candidates took this paper. Later, 37 of them went on to sit the corresponding Language B English sub-test in01.1. The paper was scheduled to last for 1h30min and was administered during a session of that length. All candidates were able to complete the paper within the period allotted and none expressed any concern about time pressure, although the results indicated this may well have been a problem.

11.1 Text

This newspaper leader text entitled "La Factura del Campo" (*ABC* 6 March 1993:17) dealt with a topic of current affairs. At 497 words the text was more than 100 words shorter than the specification. It was not edited in any way. The version used in the examination paper was a photocopy of the original onto which capital letters had been pasted to identify the paragraphs. A word-processed version appears in Table 11—1.

11.2 Task types

11.2.1 *MATCHING EXERCISE*

The first task on the paper consisted of five items. Each of these corresponded to one of the five paragraphs of the text. Essentially this was a pre-reading task, for which candidates were expected to use the target skill of scanning in order to identify the one paragraph that was referred to by each of the target phrases presented as prompts. These phrases were designed to summarize the contents of each of the paragraphs. There were five items and no distractors, meaning that a first error would lead to the loss of two points, although subsequent errors would only be penalized by one point.

Table 11—1 *La Factura Del Campo* (*ABC* 6 March 1993:17)

A NUNCA había sido escenario la capital de España de una manifestación de decenas de miles de agricultores venidos desde todos los rincones que expresara de forma más nítida y pacífica el desacuerdo de un número tan grande de españoles con la política que un Gobierno haya podido seguir para ellos.

B También es de señalar, como rasgo muy principal de la circunstancias que han rodeado esta expresión de desacuerdo hecha por el campo español, la contumacia y el sectarismo de que ha hecho gala, a lo largo de los últimos días, el servicio de propaganda del Gobierno socialista a través del principal de sus medios de actuación. Nos referimos a la manera con que los llamados Servicios Informativos de TVE han silenciado durante los pasados días la voz de los dirigentes de ASAJA (Asociación Agraria de Jóvenes Agricultores), que representa a la mayoría inmensa de los agricultores que venían hacia Madrid; mientras la palabra, la imagen y la eventual capitalización política del evento eran reservadas a los representantes de la organizaciones minoritarias de agricultores que pueden encontrarse en mayor proximidad o sintonía con el Gobierno socialista.

C De entre los discursos pronunciados por los líderes del campo ayer en Madrid destaca, como es lógico, el del presidente de ASAJA, Pedro Barato, al pedir la solidaridad nacional con los agricultores y ganaderos. Su conclusión de que Felipe González no presta la atención suficiente al campo podría merecer, sin embargo, algunas matizaciones, en el sentido de que lo que han hecho los Gobiernos socialistas ha sido atender únicamente al campo a efectos electoralistas. Las rentas agrarias y la productividad de la cabaña no han tenido peso en el quehacer de estos Gobiernos; pero sí han merecido la atención de éstos cuando por las vías espúreas de determinados subsidios destinados a paliar los efectos del paro, han sido obtenidos del campo español, especialmente en Andalucía y en Extremadura, rendimientos en forma de sufragios, sin los cuales, muy posiblemente, se habrían acortado los años de permanencia en el poder del partido de Felipe González.

D Es de prever que el impacto de la enorme y pacífica manifestación campesina de ayer tenga sus repercusiones menos en un acuse de recibo y una consecuente rectificación de la política del Gobierno –puesto que para ello, prácticamente se han agotado los plazos- que en los resultados de las próximas elecciones generales a Cortes. La magnitud que alcanza el sentimiento de agravio en el campo español permite entender que será mayor y que excederá a la fuerza y al peso de ese voto inercial, propio de la España no urbana, en la que el PSOE encontró sus reservas de votos para las dos últimas legislaturas.

E Posiblemente, en fin, no podrá separarse del resultado de las elecciones que vengan al impacto nacional logrado en Madrid por la expresión del descontento campesino. Lo hecho con el campo por los Gobiernos de González ha sido un enorme error nacional y una tremenda imprudencia política.

Table 11—2 Sub-test es01.1 Items #1-5 Matching exercise with key responses in bold italics

Lee el texto y señala a cual de los cinco párrafos corresponde cada una de las siguientes frases.

1	Las consecuencias inmediatas	*E*
2	La manifestación	*A*
3	La actitud del gobierno hacia el campo	*C*
4	El papel del Telediario	*B*
5	Las próximas elecciones	*D*

Table 11—3 Sub-test es01.1 Items #1-5 Target skills (Munby 1978:126-131)

Item	*Skill(s)*
#1-5	Scanning to locate specifically required information on a whole topic (Skill 46.5)

11.2.1.1 *Facility values*

In this case a significant number of candidates (>70%) inverted the KEY RESPONSES for Paragraphs D and E, thus failing to score on both item #1 and #5.

These introductory items were designed to be "candidate-friendly", because we wanted them to facilitate candidates' approach to the paper. Our target FVs for all of them were in the range 0.80 to 1.00, but item performance was uneven as the Figure 11—1 shows. Three out of the five items — #2, #3, and #4 — lived up to our expectations, but the other two fell far short.

The fact that items #1 and #5 produced FV scores of only 0.18 and 0.25 respectively led us to review the responses predicted as correct. For item #1, the majority of candidates had chosen Paragraph D, whereas we had predicted that Paragraph E was the correct response. On re-reading the text we decided that this item was defective because both responses were in fact correct. However, in the case of item #5, we maintained the original key response. This meant that the total score for this section was reduced to a maximum of 4, but for those who made a mistake this was really 3.

11.2.1.2 *Discrimination indices*

The clear corollary of setting a high FV target is that items are unlikely to discriminate efficiently between candidates. Our target DI score for all items was ≈0.3, and we obviously hoped that no items would produce negative DI values as this would mean that the better candidates were making errors that the weaker ones were not.

Figure 11—1 Sub-test es01.1 Items #1-5 Facility values

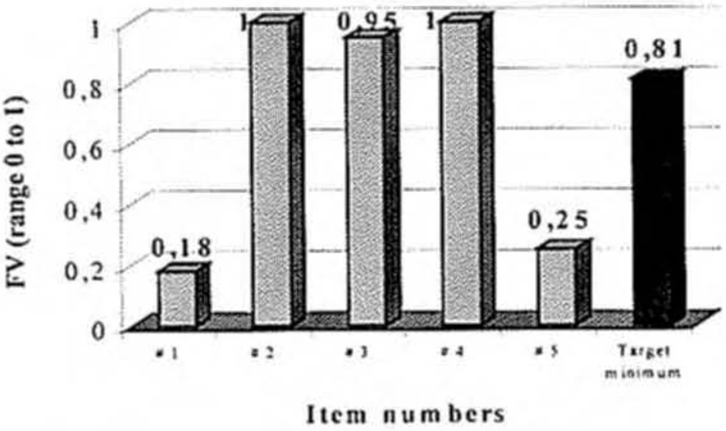
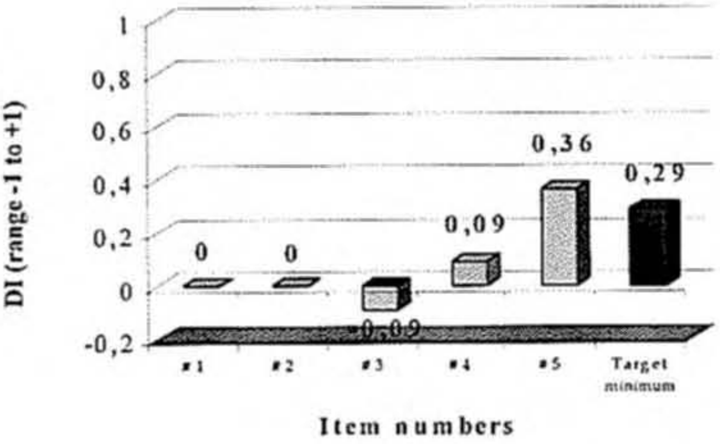


Figure 11—2 Sub-test es01.1 Items #1-5 Discrimination indices



In fact, as Figure 11—2 demonstrates, item #5 (0.36) was the only one that produced a DI score above our target minimum. All of the others failed to discriminate between candidates, and item #3 actually produced a negative value. Subsequent analysis of the scripts failed to reveal any pattern to the incorrect responses.

11.2.1.3 *Discussion*

The matching exercise did not function well. We concluded that two aspects of the task needed to be remedied. Firstly, the concept of matching a phrase to a paragraph needed to be clarified in some way or other. We decided to expand the rubric to this section of the sub-test, and make the task more specific. And secondly, we were faced with the issue of the "double penalty" — a product of the absence of any distractors leading to the inevitable loss of two points by a candidate making one mistake — which also needed to be resolved. We decided that this was best done either by reducing the number of items to four, or by adding an extra paragraph title. The former approach would have repercussions elsewhere, as we would clearly need to add an extra item to another section, because the text itself could not be "stretched" to include another paragraph. The latter would perhaps be the easier solution.

Only one of the five matching items discriminated between candidates and it was clear that in any subsequent revision of the sub-test, item #5 (FV 0.25; DI 0.36) should be retained as it stood. However, this effectively meant that the other items served no purpose other than that of providing candidates with an easy beginning to the sub-test, something that our revision would need to take into account.

11.2.2 *TRUE/FALSE/DON'T KNO W*

This test task is a three-option multiple-choice exercise. Instead of the more familiar A, B, and C options, we use "True" (T), "False" (F), or "Don't know" (?). For the candidate who carries out the task as intended, it means dealing with elements of text content. For candidates who are unable to answer the item from their knowledge of the language and who have been "exam prepared", there is a one in three chance of choosing the correct option.

This random factor, as we discussed earlier, clearly influences the validity and reliability of the items. The construct validity is affected because it is not certain whether or not the candidate actually reads the text in an effort to respond correctly. The reliability of the item is affected by the introduction of the random chance factor, which means that the part of the observed score that corresponds to elements other than language ability is increased.

At the time of the trial, we knew that candidates were not familiar with the task type and consequently it was explained to them orally when the papers were distributed at the beginning of the session. The explanation was made in English, as this was the language of instruction for all of the classes and students were accustomed to this approach. However, no example was given which, with hindsight, was seen to be an error. As a number of candidates asked for further clarification of the "Don't know" option while they were taking the paper, we concluded that this aspect had not been sufficiently clear. Three different skills were the targets of these items, as shown in Tables 11—4 and 11—5.

Table 11—4 Sub-test es01.1 Items #6-9 Target skills (Munby 1978:126-131)

Items	*Target skills*
#6-7	Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)
#8	Understanding relations between parts of text through the grammatical cohesion device of reference (Skill 32.1)
#9	Understanding explicitly stated information (Skill 20)

Table 11—5 Sub-test es01.1 Items #6-9 True/False/Don't know (the key responses are in bold italics)

Verdadero/falso/no se sabe

6	Es muy significativo que la manifestación haya tenido lugar en Madrid	***?***
7	Pedro Barato opina que los socialistas perderán las próximas elecciones	***F***
8	Los servicios informativos de TVE siempre se muestran parciales	***T***
9	Si no fuera por el voto del campo el PSOE habría perdido las últimas elecciones	***?***

11.2.2.1 *Facility values*

These items were designed to represent a challenge to candidates after the easier first task. Our target FV scores were in the range 0.30 to 0.70. And, as Figure 11—3 shows, the key responses to three out of the four items fell within this range.

However, our analysis of the actual "correct" responses supported the impression received during the exam session that candidates were not sure about the test task mechanism. Many did not choose the key response for the "Don't know" option in item #6 and, more notably, in item #9.

Figure 11—3 Sub-test es01.1Items #6-9 Facility values of key responses

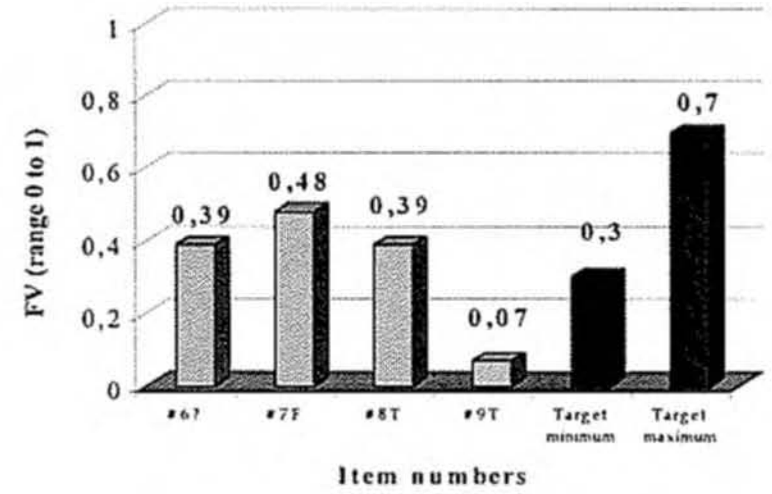
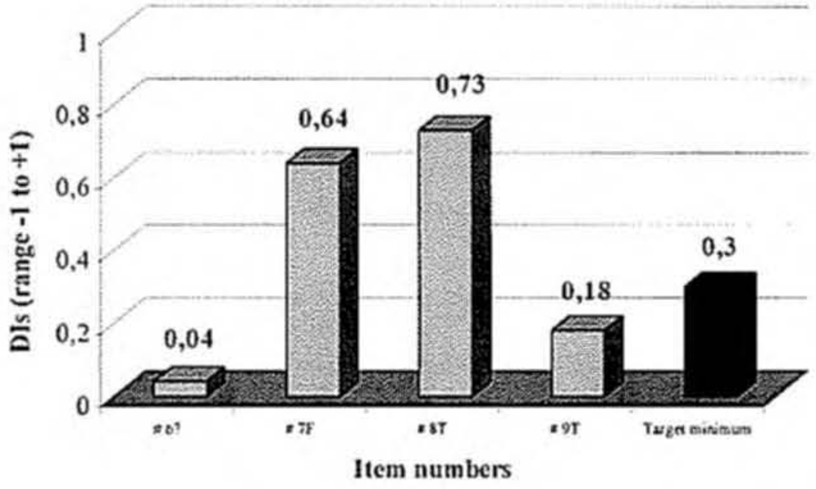


Figure 11—4 Sub-test es01.1 Items #6-9 Discrimination indices of key responses



For item #6 — *Es muy significativo que la manifestación haya tenido lugar en Madrid* —just over half of the candidates chose "T" as the correct response.

We think that this was because they did not understand that the "Don't know" option meant "there is no specific evidence for this in the text", in line with the target skill of understanding *through inference*. We conclude that the candidates did not coincide with our "correct" response because they were unfamiliar with the item type, and not because they were unable to demonstrate the ability to use the target skill.

This pattern was even more noticeable in the case of item #9, for which 93% chose "True". This option represents the more obvious inference to be drawn from the text, but is not stated explicitly.

11.2.2.2 *Discrimination indices*

The scores derived from the key responses were in line with our view that this item type was causing candidates difficulties (Figure 11—4). However, the two items that candidates responded to "correctly" produced very high DI scores. Items #7 and #8 scored 0.64 and 0.73, respectively, well above our target minimum of 0.30. Each of these items focussed on a different skill, and both proved extremely good measures of discrimination. The reason that candidates responded "correctly" to these items may well lie in the fact that the "Don't know" option was only a distractor, although the volume of responses option #7? received does seem to suggest that it was more than this. Perhaps these items were well written.

11.2.2.3 *Distractor evaluation*

When we look at the percentage of responses that each distractor attracts we are hoping to find that responses are shared equally between these. In these three-option items, we wanted each option to attract between 15% and 35% of responses.

Table 11—6 Sub-test es01.1 Items #6-9 Facility values of three-option items with the key responses indicated by shading

Options	*#6T*	*#6F*	*#6?*
FV	0.52	0.07	0.39
Options	*#7T*	*#7F*	*#7?*
FV	0.07	0.48	0.45
Options	*#8T*	*#8F*	*#8?*
FV	0.39	0.48	0.14
Options	*#9T*	*#9F*	*#9?*
FV	0.93	0.00	0.07

Table 11-6 shows us that only item #8 produced a balance of responses approaching our target, with the lowest percentage of incorrect answers (14%) going to option #8? With all of the other items the options performed unevenly. For example, in the case of item #7, based on the same skill as #6 and with "False" (#7F) the key response, a large percentage of candidates answered correctly (48%), but the incorrect responses were shared unequally between the distractors.

11.2.2.4 *Discussion*

This item type called into question the FACE VALIDITY of the sub-test. Candidates later reported informally that this item type was difficult in terms of the concept implied by the third option "Don't know". We considered three possible solutions: adding an example item, and one or two easier items as the first in the series, and writing a more detailed rubric. We also debated the idea of abandoning the task type completely. However, in view of the fact that two items had produced very high DI values we decided that in the first instance the task type needed perfecting rather than discarding.

These items did little to help us in our search for close ties between task types and target skills. The two items based on the same skill — #6 and #7 — produced similar results in terms of facility, but not for discrimination. The only potential link that we thought we might be able to exploit was that between Skill 32.1 and the T/F/? format, as item #8 scored within our target range for facility and was a good discriminator.

ITEM #8

The prompt of item #8 was:

> *Los servicios informativos de TVE siempre se muestran parciales*

And the key response was to be found in the middle of Paragraph B:

> ***B*** *...la contumacia y el sectarismo de que ha hecho gala, a lo largo de los últimos días, el servicio de propaganda del Gobierno socialista a través del principal de sus medios de actuación. Nos referimos a la manera con que los llamados Servicios Informativos de TVE han silenciado durante los pasados días...*

The key to finding the correct response lies in the reader's ability to establish the cohesive reference between the second sentence, in which *TVE* is mentioned directly, and the previous phrase, in which the expression *el servicio de propaganda del Gobierno socialista* is the referent.

With a view to revising and refining these items we decided that we could maintain #6, #7, and #8, but that #9 would require revision. We also felt that if it were possible to add a further item targetted on skill 32.1, as #8 was, this would possible give us a valuable addition to the sub-test.

11.2.3 *SENTENCE COMPLETION ITEMS*

Six sentence completion items were included in this sub-test, all of which are reproduced in the table below. The task type was sufficiently well-known to candidates that we believed they would have no difficulty in responding adequately. However, the reality of candidate responses was such that it was extremely difficult to apply a dichotomous (right-wrong) score to most of the candidates' responses. We had already specified that the test should involve the maximum objectivity in marking, and the responses that appeared made this aim impossible to achieve.

Table 11—7 Sub-test es01.1 Items #10-15 Question stems

Completa las siguientes frases de la manera más apropiada.

10. ______________________________ requiere algunas matizaciones.
11. ______________________________ más importantes de la manifestación se
12. Prevén en ______________________________.
13. No podrán separarse ______________________________ de ______________________________.
14. Como resultado de la presentación de la noticia por ______________________________.
15. Una rectificación por parte del gobierno ______________________________.

Many of these threw up a number of difficulties that affected our concerns for the maximum objectivity of marking in order to avoid difficulties of inter-rater and intra-rater reliability. The responses given raised questions such as "Is a response correct if it includes a spelling or grammar mistake?" This type of open answer requires a detailed markscheme and an investment in rater training and moderation. Both of these were beyond the control of the test designers aiming, as we were to produce a test that would be highly objective. In conclusion, we chose to discard this task type and did not carry out any statistical analysis of performance.

11.2.4 *FIVE-OPTION MULTIPLE CHOICE QUESTIONS*

There were a total of eight multiple-choice questions (MCQ) in the sub-test and these divided into two groups. Items #16-19 focussed on aspects of comprehension, while the other four targeted vocabulary.

Table 11—8 Sub-test es01.1 Items #16-23 Target skills (Munby 1978:126-131)

Item	*Target skill*
#16, #18, #19	Understanding explicitly stated information (Skill 20)
#17	Understanding the communicative value (function) of sentences and utterances with explicit indicators (Skill 26.1)
#20-23	Deducing the meaning and use of unfamiliar lexical items, through understanding word formation (Skill 19.1) and context (Skill 19.2)

The purpose of using five option items was to reduce the "guess factor" to 20% and increase the reliability of the items. With five options to choose from a candidate who only answered items at random would be likely to score 1 in 5 right. With a three- or four-option item the chances are 1 in 3, and 1 in 4, respectively. The five option items are, however, much more difficult to write as we have mentioned earlier.

11.3 Conclusions

Given the wholly experimental nature of a trial exercise like this, little could be expected in terms of reliability, and the coefficient recorded was actually negative, indicating extremely poor performance. The key factors leading to this included the very few items involved and the relatively small population on which it was trialled. In the light of this, there is no value in discussing the Standard Error of Measurement, nor the descriptive statistics.

As a result of our analysis of this trial paper, and of sub-test in01.1, we made a number of decisions which we incorporated into the revised version of this sub-test and which formed the basis of es01.2, es02, and in02.

Among other things, we decided

- To extend the time allowed to 2 hours
- To reduce the range of task types to three, maintaining only those that can be objectively marked

- To concentrate on the skill and test task combinations that produced acceptable FV and DI scores. Specifically, the combination of skills 22.1 — *Understanding relation between parts of text through the grammatical cohesion device of reference* — and 22.1 — *Understanding information in the test, not explicitly stated, through making inferences* — with the True/False/Don't know task type.

- To review the FACE VALIDITY of the True/False/Don't know items.

12 SUB-TEST IN01.1

12 SUB-TEST IN01.1

12.1 TEXT
12.2 TASK TYPES
12.2.1 *Matching exercise*
12.2.1.1 Facility values
12.2.1.2 Discrimination indices
12.2.1.3 Discussion
12.2.2 *True/False/Don't know*
12.2.2.1 Facility values
12.2.2.2 Discrimination indices
12.2.2.3 Discussion
12.2.3 *Five-option multiple choice questions*
12.2.3.1 Discussion
Item #11
Items #12, #13, and #14
Items #15 and #16
12.3 RELIABILITY
12.4 DESCRIPTIVE STATISTICS
12.5 CONCLUSIONS

THIS SUB-TEST represents the first foreign language paper that we prepared within this research study. We believe it is more typical of the type of test candidates would be accustomed to. The application of foreign language test methods to Language A, which we have seen in es01.1 would be something unusual for them, and might have had some influence on performance. Despite the fact that some though not all of the test tasks were familiar, we always need to make allowances for the unexpected. As Bachman (1990) suggests, the test-taker is central to the test-taking process, and any factor that may influence test-taker behavior and therefore contribute to the observed score, is obviously important.

In a class session after taking sub-test es01.1, 37 of the original 44 students from the first year *Traducción I (Inglés)* course sat sub-test in01.1. We are not able to explain the reduction in the number of candidates other than by saying that this was probably due to the vagaries of student attendance patterns. It was certainly characteristic of this particular group of students. Candidates had not been told there was to be a second test in order to ensure that a statistically sound number (≥30) attended. From previous experience, we felt that asking students to sit an English test for no apparent purpose would have produced an unnatural group composition and could easily have reduced the number attending below the required minimum. We hoped that this would ensure no one prepared for the test, and that no one stayed away from class because of it.

As with sub-test es01.1, we presented this paper to candidates as a reading comprehension activity similar to other work they had been doing. In a follow-up session, when we gave candidates their marks, we told them that the tests had something to do with preparations for the aptitude test for entry to the new degree program.

In our analysis of the results, as in the case of sub-test es01.1, we need to bear in mind the fact that the number of candidates, who sat the paper, while statistically sound, was nonetheless limited. This means that when we draw our conclusions we must do so with a measure of skepticism. The smaller the sample, the less certain we can be about any possible generalizations.

12.1 Text

The text "Short's breakaway may be a blunder" (*The Independent on Sunday*, 28 February 1993; Home:5) was unedited and the version presented was a photocopy onto which letters had been pasted to identify the paragraphs. We reproduce a word-processed version of the original text in Table 12—2. The text met the test specification at 396 words. The topic was of current affairs, and it did not require any specialist knowledge. We

Table 12—2 "Short's breakaway may be a blunder" (*The Independent on Sunday*, 28 February 1993; Home:5)

A LEADING figures in the British chess world are concerned that Nigel Short may have made a tactical error in backing a breakaway organisation which is attempting to seize control of the world championship.

B They fear that the Professional Chess Association, whose formation was hastily announced on Friday, still has no constitution and could therefore favour the cause of Garry Kasparov, the defending champion, because no safeguards exist to protect the challenger.

C There was also concern that the joint statement by Short and Kasparov that they were refusing to play their title match in Manchester under the auspices of Fide, the World Chess Federation, might mean that the championship would eventually take place outside Britain.

D It emerged yesterday that the new chess body's lack of a constitution means that it has no mechanism for resolving disputes such as the choice of venue for the championship.

E Raymond Keene, chess correspondent of the Times, who helped draft the announcement of the new organisation, said that the players would be working together to evolve a workable constitution.

F Murray Chandler, a British grandmaster since 1983 and editor of British Chess magazine, said that Kasparov, as the champion, was in a stronger position than Short because there would be no independent authority to force him to defend his title.

By Andrew Gliniecki

G Short might have rushed into the decision because of is anger at not being consulted about Fide's decision to choose Manchester as a venue, he said.

H Mr Chandler went on: "Kasparov has a strong personality, and without Fide there is no independent body to make sure the challenger gets a fair crack of the whip."

I He said there was a widespread dissatisfaction among British players but the situation required a considered response. "I'm not saying that an alternative to Fide isn't preferable, but it's got to be the right alternative. The history of chess is littered with new organisations which come out of nowhere then quickly disappear."

J Simon Brown, international director of the British Chess Federation, said that he was "extremely nervous" about the new body and hoped that a compromise could yet be hammered out between the players and Fide.

K He added: "This new body is apparently to invite new bids for venues, so the championship could yet be played outside Britain."

consulted members of the Examination board and they agreed that the topic was unlikely to favor any particular group of candidates, although a few of them did criticize the choice text in informal feedback. They commented that followers of the world chess scene would perhaps have had an advantage. Given the very specific nature of the article, we do not consider that this was actually the case, but as the text did raise the issue of FACE VALIDITY, we cannot ignore it completely.

12.2 Task types

12.2.1 *MATCHING EXERCISE*

The text consisted of 12 paragraphs, identified by the letters A to K, and five matching items were included in the test. This meant that seven of the paragraphs served as distractors. Consequently, an advantage of this paper over sub-test es01.1 was that we would not penalize any candidate twice for a single error, as was the case on the LA paper. In analyzing the scripts we found only two candidates out of the 288, had made an error possibly induced by the task type.

Table 12—1 Items #1-5 Target skills (Munby 1978:126-131)

Item	*Target skills*
#1-5	Scanning to locate specifically required information on a whole topic (Skill 46.5)

Items #1 to #5 aimed to test candidates' ability to scan the text, and all of the items focused on the same skill, as was the case in sub-test es01.1. In reality, in order to candidates to demonstrate their ability to use the skill in question they also demonstrate they can identify the important information present. The nature of the task is such that recognition is the key to successfully answering each item.

12.2.1.1 *Facility values*

The range of FVs was high, as predicted, with all of the items passing the target minimum. Items #1 and #5 both scored 0.84, which is sufficient to achieve our global aim for these items while at the same time demonstrating a minimum level of difficulty. The others were all much closer to 1.00, indicating that they were very easy.

12.2.1.2 *Discrimination indices*

DIs for three of these items — #3, #4, and #5 — were higher than might have been expected, and coincided with the pattern of facility values. Items #1 and #5 produced comparatively better

Figure 12—1 Sub-test in01.1 Items #1-5 Facility values

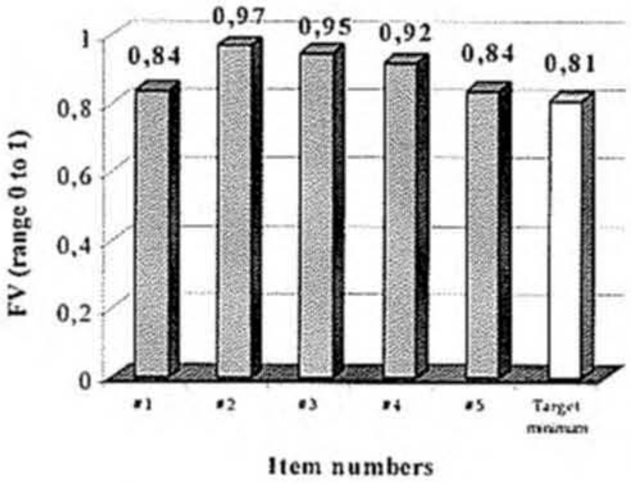
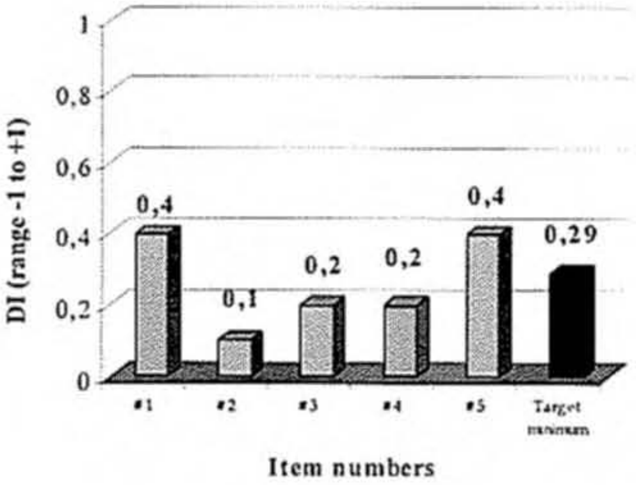


Figure 12—2 Sub-test in01.1 Items #1-5 Discrimination indices



FV scores and discriminated well. Items #3 and #4 were both very easy, and barely discriminated between candidates.

12.2.1.3 *Discussion*

Table 12—3 Items #1-5 Question stems with the key responses in bold italics

Scan the text and identify which paragraphs are summarised by the five sentences below.

1	Short's motives	G
2	A joint statement	C
3	Help from a journalist	E
4	Kasparov's personality	H
5	Other players' reactions	I

In general terms the performance of these items was somewhat better than might have been expected. The combined FV and DI scores produced by items #1 and #5 are particularly interesting. The items, which appear in Table 12—3, were straightforward and in most cases we believed that candidates would be able to respond from scanning the text without the need to read, or even re-read, more extensively. We thought that items #2 and #4 were the easiest as the words appearing in the question stems also appeared in the text itself. Item #2 indeed proved that easy.

Item #3 required candidates to understand that "chess correspondent of the *Times*" — the target phrase in the text — was linked to the synonym "journalist", which appeared in the stem. They also needed to see that the verb form "helped", in Paragraph E, echoed the noun "help" in the stem.

Both item #1 and #5, were, we thought, more difficult. In #1 candidates needed to achieve two tasks. Firstly they had to identify the appropriate paragraph, which they could initially do by scanning the text to locate the name "Short". This would lead them to identify Paragraphs A, C, F, and G, at least, as the potential sources of the key response. Secondly they would need to identify within each of these paragraphs the presence of lexical items associated with "motives" or motivation, the second element of the stem. In this case the correct answer would be found by interpreting the phrase

> *Short might have rushed into this because of his anger...*

as describing supposed motives.

Finally we come to Item #5, which we thought easier than #1, but which did, in fact provide the same statistical response. The stem links with the topic sentence of Paragraph I:

> *He said there was widespread dissatisfaction among British players...*

The stem and the target paragraph both include the noun "players", and the interpretative element lies in linking the noun "reactions", in the stem, to "dissatisfaction", in the text. As we have already commented, this item also produced the best figures.

12.2.2 *TRUE/FALSE/DON'T KNO W*

As all of the candidates had recently taken test es01.1, we knew that they had acquired experience of the True/False/Don't know task type. Consequently, we set greater store by the results achieved by this task type on this paper than by those from es01.1.

This section contained only three items, however the overall performance was promising with regard to candidate perception of the item type. In the sub-test taken previously, this task type had been new. As we have described above, candidates were uncertain as to the meaning of the "Don't know" option. In fact, we identified only one candidate who may have made an error in handling the task type. The fact that item #6, for which "?" was the key response attained a FV of 0.62 seems promising. The task type may well have been "learned" and candidates are therefore performing within their capabilities. This would indicate that the item itself is also performing well, and that it may serve a useful purpose in the final test paper specification.

Table 12—4 Items #6-8 Question stems with the key responses in bold italics

True/false/questionable

6	Manchester may be the venue chosen to host the world title match	*?*
7	The new organisation is called Fide	*F*
8	Short is the favourite to win the title	*?*

Table 12—5 Target skills (Munby 1978:126-131)

Item	*Target skills*
#6, #8	Understanding information in the text, not explicitly stated, through making inferences (Skill #22.1)
#7	Understanding explicitly stated information (Skill #20)

12.2.2.1 *Facility values*

If we look at Table 12—6, we see that both items #7 and #8 reached FV scores slightly above and at our target at 0.78 and 0.70 respectively. Moreover, item #8 also provides valuable data on the test task in that the "?" option attracted 27% of responses. If we link this to the skill in question, namely inferencing, we see that here candidates had read the text carefully and reached their decision on the basis of a number of words and phrases that appear throughout the text. All of these lead to the conclusion that Nigel Short is far from being the favorite. A conclusion that directly contradicts the proposition put forward in the question stem.

Table 12-6 Facility values of three-option items (Shading indicates key responses)

Options	*6T*	*6F*	*6?*	***Target range***
FV	0.16	0.22	0.62	0.3-0.7
Options	*7T*	*7F*	*7?*	
FV	0.22	0.78	0.00	
Options	*8T*	*8F*	*8?*	
FV	0.05	0.70	0.27	

12.2.2.2 *Discrimination indices*

These items all give very good DI scores, considerably higher than our minimum, as Table 12—7 shows.

The data support the view that the stronger candidates had managed to come to terms with the task type. If we consider these in combination with the FV scores we can assume that some of the less able candidates had done so too: otherwise the FV scores would have been rather lower.

12.2.2.3 *Discussion*

With the proviso that we are only dealing with three items, we can say that in this paper the task type is both manageable — from the candidates' point of view — and apparently serves a purpose as an instrument of measurement. What's more, this would particularly seem to be the case when we are dealing with the target skills of understanding information both explicit and implicit in the text.

12.2.3 *FIVE-OPTION MULTIPLE CHOICE QUESTIONS*

Sub-test in01.1 included six MCQ items that focussed on a range of skills. By contrast with es01.1, we identified the parts of the text to which these items referred in the question stems, such as that presented in Table 12—7. We did this to test out the theory that it would facilitate candidates' progress through the foreign language text — something that we did not think necessary for LA sub-test es01.1.

12.2.3.1 *Discussion*

In order to illustrate the performance of these items we now discuss the most important of them in turn.

ITEM #11

Table 12—7 Item #11 Question stem and options

11 The breakaway organisation (para A) ____
A Has been established by Raymond Keene ____
B Has chosen a new venue for the match ____
C is supported by Garry Kasparov ____
D Was set up by Kasparov and Short ✓
E is called the British Chess Federation ____

This item is perhaps the most interesting of the set. The task proved difficult, with a very low FV of 0.22, below our target, and a good DI of 0.4. If we compare the responses to all of the options, as set out in Table 12—8, we see that option #11B obtained the highest percentage. We must ask ourselves, therefore, whether or not this was the "correct" response, rather than the key response we had chosen at the time of writing the paper.

Table 12—8 Item #11 Facility values for all options (shading indicates key response)

Item	*#11A*	*#11B*	*#11C*	*#11D*	*#11E*
Actual responses	0	15	9	8	2
FV	0.00	0.41	0.24	0.22	0.05
Total responses	34				
Errors	3				
Errors/Total(%)	8.1%				

A re-reading of the text makes it clear that #11B cannot be correct. The information given is that no venue has been chosen for the event, and that there is still much doubt over whether or not the championship will take place in Britain or elsewhere. This is highlighted in the phrases

...might mean that the championship would eventually take place outside Britain." (Paragraph C)

and

"...so the championship could yet be played outside Britain." (Paragraph K)

This item, then, is more difficult than was out intention, but it discriminates well, with a DI score of 0.4 (Figure 12—3).

It is worth noting that the last three rows of Table 12—8 give the figures for candidate responses and errors. These show that although 37 candidates took the paper, 3 of them failed to

Figure 12—3 Sub-test in01.1 Items #11-16 Discrimination indices

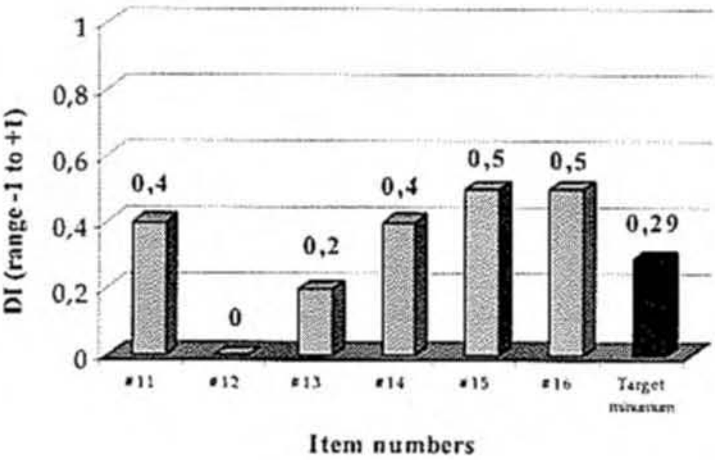
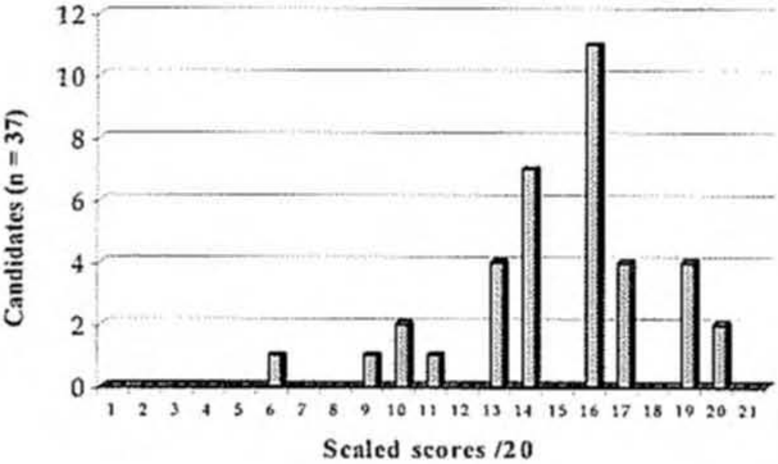


Figure 12—4 Sub-test in01.1 Frequency distribution of scaled scores



respond adequately to this item. Failing to respond adequately may mean marking two or more options as correct, or not clearly marking any of them. The figure in this case is low enough for us to discount it.

ITEMS #12, #13, AND #14

Item #12 is very easy and has a DI score of 0, and item #13 is also above our target FV, and therefore very easy. It discriminates to a limited extent only. Item #14 achieves a FV within our target range, and achieves minimal discrimination, too.

ITEMS #15 AND #16

The remaining two items #15 and #16, are worthy of further comment. Both achieve FV scores within our range, and very good DI values. Both of these items focus on the same skill, namely that of "reading between the lines". Moreover, in the case of #16, there is some spread of responses among the four distractors, although #16B attracted none at all. A fresh look at the text does not provide us with any clear indication as to why none of the candidates chose this response. It was included in the set of options because it reflects part of the overall content of the text, and echoes the phrase "new organisations", which comes near to the end of the paragraph (Para. I in Table 12—2).

Table 12—9 Item #16 Facility values for all options (shading indicates key response)

Item	*#16A*	*#16B*	*#16C*	*#16D*	*#16E*
Actual responses	19	0	6	3	9
FV	0.51	0.00	0.16	0.08	0.24

Table 12-10 Items #15 and #16 Question stems and options (a ✓ indicates the key response)

15 The sentence "Kasparov has a strong personality ... of the whip" (para H) is intended as
A an assessment
B a warning ✓
C a new insight
D a joke
E a forlorn hope
16 The sentence "I'm not saying that ... then quickly disappear" (para I) serves as
A a balanced qualification ✓
B a new alternative
C the definitive response
D the last chance
E a tried and tested response

Among these MCQ items, candidate response to distractors ranged from 0.00 to 0.78, with only nine out of forty attracting >15% of responses. These figures indicate that the options would need very careful revision in order to achieve a more balanced spread.

12.3 Reliability

Table 12-11 Reliability coefficient

Kuder Richardson KR20	**0.357**

The very limited number of items in this sub-test makes it inevitable that any coefficient of reliability will be low. We calculated the KUDER-RICHARDSON KR-20 formula, and recorded a coefficient of 0.357, which is completely unacceptable. However, when we applied the SPEARMAN-BROWN PROPHECY FORMULA, we found that by almost doubling the number of items we could obtain a coefficient of 0.928. This would, of course, mean that the extra items should be as similar as possible to those currently in the test. While this is theoretically possible, the practical question we would need to answer is does the text actually provide content sufficient for the writing of 13 more items? Our initial impression is that this is highly unlikely. Another six items, taking the sub-test to 20, would achieve a coefficient of 0.678. For one sub-test, this would not be satisfactory, but in the context of a set of sub-tests it might be acceptable.

Table 12—12 How many items are needed to achieve a satisfactory level of reliability?

Spearman-Brown Prophecy Formula (27 items)	**0.928**

The STANDARD ERROR OF MEASUREMENT (S_e) for this sub-test is 1.65. This means that we can be 68% certain that any individual candidate's true score will be in the range of the observed score ±1.65. While this appears a small quantity, we must bear in mind the fact that the sub-test only adds up to 14 raw points. The margin then, in percentage terms, is quite significant. This, then, is another factor that leads us to question the value of our results.

Table 12—13 How "true" are candidates scores?

S_e	**1.65**

12.4 Descriptive statist ics

The descriptive statistics calculated for the raw scores of candidates, with totals out of 14, appear in Table 12—12. In addition, the frequency distribution of scores scaled out of a total of 20 to enable easy comparison with other sub-tests in the study, is depicted in Figure 12—4.

Table 12—13 Descriptive statistics

Mean	9.46
Median	10.00
Mode	10.00
Standard error	0.35
Standard deviation	2.14
Skewness	(0.89)
Variance	4.59
Level of confidence : 95%	0.69

The three measures of centerdness, mean, median, and mode, are all very close demonstrating that the distribution is quite "normal". The deviation from normal is shown in the figure of –0.89 for SKEWNESS, which describes the curve of the tail to the left of the center. This negative skew is associated with an easy test.

12.5 Conclusions

In the trialling of this sub-test, we have found that even matching exercise items that achieve high FV scores can discriminate well between candidates, as demonstrated by items #1 and #5. Both of these items involved the use of the target skill of

> *Scanning to locate specifically required information on a whole topic (Skill 46.5)*

In each case the depth of search involved was probably quite complex.

Furthermore, as in the case of the True/False/Don't know items, we have seen that once candidates have "learned" the task type, these items perform reasonably well. The skills that have proved successful in this instance have been

> *Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1*

and

> *Understanding explicitly stated information (Skill 20).*

In the case of the former skills, this supports the results of sub-test es01.1.

The MCQ items, with five options, have shown us the difficulties involved in writing four balanced distractors, although some of the items have proved successful. In particular, we have identified items #15 and #16, which require candidates to "read between the lines" to be valuable parts of the sub-test. This is

> *Interpreting text by going outside it: "reading between the lines" (Skill 34.2).*

In the light of these factors, and those we have described in our analysis of sub-test es01.1, we revised es01, and wrote sub-tests es02 and in02, for use with Cohort A.

13 SUB-TEST ES01.2

13 SUB-TEST ES01.2

13.1 TEXT
13.2 TASK TYPES
13.2.1 *Matching exercise*
13.2.1.1 Facility values
13.2.1.2 Discrimination indices
13.2.2 *True/False/Don't know*
13.2.2.1 Facility values
13.2.2.2 Discrimination indices
13.2.3 *Discussion*
13.2.4 *Five-option multiple choice questions*
13.2.4.1 Facility values
13.2.4.2 Discrimination indices
13.2.4.3 Distractor evaluation
13.2.4.4 Discussion
13.3 DESCRIPTIVE STATISTICS
13.4 RELIABILITY
13.5 CONCLUSIONS

THIS TEST was a revised version of es01.1, the first Spanish language test we trialled. Modifications to the test were carried out in the light of the analysis of the trial version, and of the trial English language sub-test, in01. The final version was studied by a number of teachers in the Faculty in order to ensure construct validity. To make reading easier, we reproduce the text again in the table below. Items are reprinted by task type in the corresponding sections throughout the chapter, and KEY RESPONSES appear in bold italics.

Sub-test es01.2 provided us with thirteen items that achieved both FV and DI scores within our target ranges. These items could therefore provide is with material for future test design. The first four items all fell within their target range in that they were designed to be easy, and nine other items met our target. A further positive result is the fact that no item produced a negative DI score. The value of the particular combinations of task type and target skill would need to be supported, or not, by results from other tests in order for us to be confident about their value.

13.1 Text

In July 1993, 288 (64%) out of 449 candidates took this test, and the remaining 36% took the parallel test es02. The division was based firstly on the candidates' choice of Language B, and, secondly, on alphabetical order of surnames. Parallel tests in English, French, and German were administered in the same session, which was held in the Faculty of Arts building, University of Granada.

The Examination board of 11 lecturers was responsible for invigilating the examination in two separate rooms. In line with the results of the trial test, this sub-test was scheduled to last a maximum of two hours, and not 1h 30 min, to ensure candidates were able to complete all items without time pressure, and most candidates were able to complete the paper within that period.

We will now look at how each item performed by breaking down our analysis along the lines of the task types chosen, as these separated the examination into three sections.

Table 13—1 *La Factura Del Campo* (*ABC* 6 March 1993:17)

A NUNCA había sido escenario la capital de España de una manifestación de decenas de miles de agricultores venidos desde todos los rincones que expresara de forma más nítida y pacífica el desacuerdo de un número tan grande de españoles con la política que un Gobierno haya podido seguir para ellos.

B También es de señalar, como rasgo muy principal de la circunstancias que han rodeado esta expresión de desacuerdo hecha por el campo español, la contumacia y el sectarismo de que ha hecho gala, a lo largo de los últimos días, el servicio de propaganda del Gobierno socialista a través del principal de sus medios de actuación. Nos referimos a la manera con que los llamados Servicios Informativos de TVE han silenciado durante los pasados días la voz de los dirigentes de ASAJA (Asociación Agraria de Jóvenes Agricultores), que representa a la mayoría inmensa de los agricultores que venían hacia Madrid; mientras la palabra, la imagen y la eventual capitalización política del evento eran reservadas a los representantes de la organizaciones minoritarias de agricultores que pueden encontrarse en mayor proximidad o sintonía con el Gobierno socialista.

C De entre los discursos pronunciados por los líderes del campo ayer en Madrid destaca, como es lógico, el del presidente de ASAJA, Pedro Barato, al pedir la solidaridad nacional con los agricultores y ganaderos. Su conclusión de que Felipe González no presta la atención suficiente al campo podría merecer, sin embargo, algunas matizaciones, en el sentido de que lo que han hecho los Gobiernos socialistas ha sido atender únicamente al campo a efectos electoralistas. Las rentas agrarias y la productividad de la cabaña no han tenido peso en el quehacer de estos Gobiernos; pero sí han merecido la atención de éstos cuando por las vías espúreas de determinados subsidios destinados a paliar los efectos del paro, han sido obtenidos del campo español, especialmente en Andalucía y en Extremadura, rendimientos en forma de sufragios, sin los cuales, muy posiblemente, se habrían acortado los años de permanencia en el poder del partido de Felipe González.

D Es de prever que el impacto de la enorme y pacífica manifestación campesina de ayer tenga sus repercusiones menos en un acuse de recibo y una consecuente rectificación de la política del Gobierno -puesto que para ello, prácticamente se han agotado los plazos- que en los resultados de las próximas elecciones generales a Cortes. La magnitud que alcanza el sentimiento de agravio en el campo español permite entender que será mayor y que excederá a la fuerza y al peso de ese voto inercial, propio de la España no urbana, en la que el PSOE encontró sus reservas de votos para las dos últimas legislaturas.

E Posiblemente, en fin, no podrá separarse del resultado de las elecciones que vengan al impacto nacional logrado en Madrid por la expresión del descontento campesino. Lo hecho con el campo por los Gobiernos de González ha sido un enorme error nacional y una tremenda imprudencia política.

13.2 Task types

13.2.1 *MATCHING EXERCISE*

These were the first four items on this paper, a reduction from the five included in the trial test. The target skill for all four items was the same as in the trial version, but sub-test es01.1 item #1 had been cut, item #2 had been modified, and all four items were now presented with the target phrase in inverted commas, to reinforce their presentation as possible "titles" for the paragraphs. In addition, the rubric now stated clearly that answers were to be written onto the separate answer sheet.

Table 13—2 Items #1-4 Target skills (Munby 1978:126-131)

Item	*Skill(s)*
#1-4	Scanning to locate specifically required information on a whole topic (Skill 46.5)

Table 13—3 Items #1-4

Lee el texto y señala a cual de los cinco párrafos podría corresponder cada una de las siguientes frases como título. Subraya la letra en la hoja de respuestas.

1	"El acontecimiento".	*A*
2	"La actitud del gobierno hacia el campo".	*C*
3	"El papel del Telediario".	*B*
4	"Las próximas elecciones".	*D*

We decided to reduce the number of matching tasks to four because their purpose was not to differentiate between candidates, but to create a favourable impression of the sub-test. Gaining an extra item for the test "proper" was a clear bonus.

13.2.1.1 *Facility values*

As in the trial version, we wanted candidates to feel confident at the beginning of the paper that they would be able to answer the items, so these were designed to be easy. We were certain that they would produce high FVs, and that they would all be "very easy" (>0.80). In this, our plan was successful as indicated by the resulting FVs, shown in Figure 13—1.

All four items exceeded our target. Since our prediction was correct, the corollary of such high FVs was that the items would fail to discriminate between candidates, as we will later see.

Three items were maintained with only minimal modification from the trail version to the full version of the sub-test. Numbering of the items differed from trial to final version, and accordingly we have not included the item numbers here. As the table shows, in two cases the high FV scores were maintained from the earlier to the later version. In the third instance, there was a substantial increase in the score. This was due to the fact that the error candidates had led to commit with the trial sub-test, due to the two overlapping items, had been remedied by omitting one of these from the set.

Table 13—4 A comparison of facility values for items included unchanged in sub-tests es01.1 and es01.2

Es01.1	*es01.2*
0.95	0.97
1.00	0.99
0.25	0.85

13.2.1.2 *Discrimination indices*

By making the first four items easy, we had effectively reduced to 16 out of 20, the number of items that could serve to differentiate between candidates. The DI values for these four items were below our target, as Figure 13—2 shows, but were acceptable given their purpose.

13.2.2 *TRUE/FALSE/DON'T KNO W*

If we compare this section of the sub-test with the equivalent section of the trial version, we can see that there are two items more in the present version, than there were in the first. Three of the original items have been maintained unchanged; the fourth has been modified; and two new items have been added. These changes were made for a number of reasons. Firstly, we thought that the DI scores of the trial sub-test were satisfactorily high enough to justify trying to include more, similar items. We had already omitted one of the matching exercise items, and so were able to include at least one extra here.

Figure 13—1 Sub-test es01.2 Items #1-4 Facility values: these items all surpassed the (already high) target minimum of 0.81

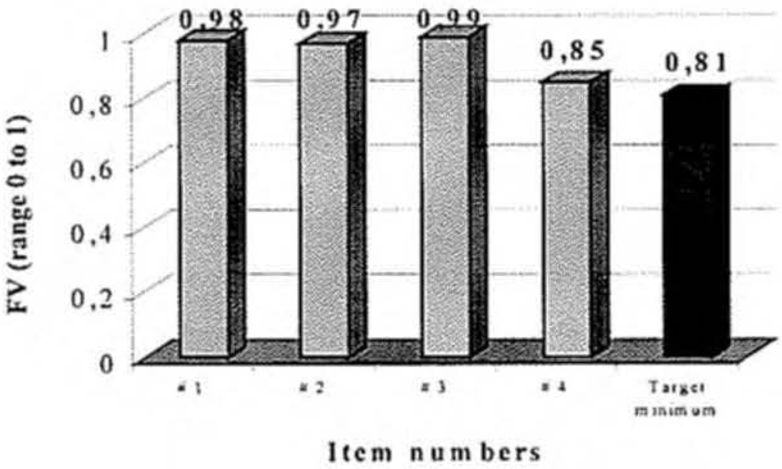
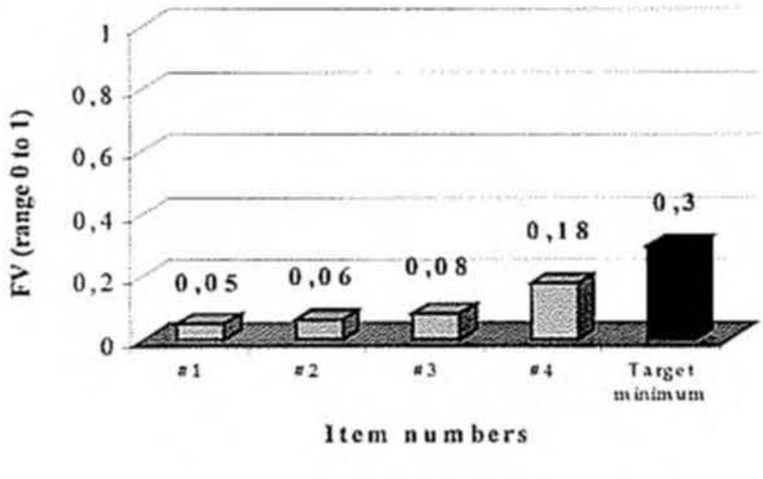


Figure 13—2 Sub-test es01.2 Items #1-4 Discrimination indices: as was to be expected, these items failed to achieve even our target minimum



In addition, we had discussed the overall format of the sub-tests with those colleagues preparing the examinations in Spanish and other languages, and collectively reached the decision that the maximum similarity of format from one sub-test to another was essential. While there were discrepancies over the relative value of this task type, it was agreed that the ratio between the matching tasks and the T/F/? Items, should be 4:6, or 5:5. The final choice was left to the discretion of the initial question-writer in each case, bearing in mind the qualities and characteristics of the specific texts being used.

One of the concerns arising from analysis of the trial sub-tests, and voiced again by some members of the Examination board had been that the task type might have been the cause of candidate errors. In view of this the first statistic of importance was the error rate, which indicated how many candidates failed to answer each item in the correct manner. These five items produced a total of three candidate errors in total, and all of these were committed by the same individual, who responded twice to each of items #7, #8, and #9. The candidate in question responded correctly both in manner and response, to items #9 and #10. From these data, we deduce that the rubric, and the time allowed were together sufficient to eliminate this source of error for 287 of the 288 candidates. For the individual concerned, we believe that the error was a problem of indecision over the options and a wish either to "hedge bets" by offering two responses, or an oversight of revision. In general, though, we conclude that the task type was not a serious source of candidate error.

In our conclusions to the results provided by es01.1, we had seen the value of including tasks focussed on the skill of

> *Understanding relations between parts of text through the grammatical cohesion device of reference (Munby 1978:128-131 Skill 32.1).*

It was our initial intention to focus at least one of the two extra items on this skill, but in practice this proved difficult, as the text did not lend itself to such items. The two extra items were accordingly focussed on a more accessible skill, namely that of

> *Understanding explicitly stated information (Munby 1978:128-131 skill 20).*

We list the target skills of all six items in Table 13—5, and the items in Table 13—6.

Table 13—5 Items #5-10 Target skills (Munby 1978:126-132)

Item(s)	*Skills*
#5-6	Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)
#7	Understanding relations between parts of text through the grammatical cohesion device of reference (Skill 32.1)
#8-10	Understanding explicitly stated information (Skill 20)

Items #5 and #6 were based on candidates' understanding of the ideas explicit or implicit in the texts. Items #8, #9 and #10 required the same skill. By contrast, #7 called on the reader to understand the relationship established through the lexical parallel of two sentences in Paragraph B. This was based on the use of Skill 32.1 as it required the understanding of the complex relationship of textual cohesion through which the author makes ironic use of co-reference.

Table 13—6 Items #5-10 Rubric and questions (key responses are in bold italics)

V (Verdadero), F (Falso) o ? (el texto no informa sobre este punto), y subraya la letra o el símbolo correspondiente en la hoja de respuestas.

5.	Es muy significativo que esta manifestación haya tenido lugar en Madrid.	***?***
6.	Pedro Barato opina que los socialistas perderán las próximas elecciones.	***F***
7.	Los servicios informativos de TVE siempre se muestran parciales.	***T***
8.	Si no fuera por el voto del campo el PSOE nunca habría llegado a gobernar.	***T***
9.	En opinión del autor de este artículo el campo seguirá apoyando al Gobierno a pesar de lo hecho.	***F***
10.	Las consecuencias más importantes de la manifestación se verán especialmente en Andalucía y Extremadura.	***?***

13.2.2.1 *Facility values*

When we came to analyze the actual "correct" responses given by candidates we found that these produced FVs which conflicted with our key responses.

Table 13—7 Items #5-10 Facility values for three-option items (shading indicates key responses)

Options	*5T*	*5F*	*5?*	***Target***
FV	0.64	0.15	0.22	0.3-0.7
Options	*6T*	*6F*	*6?*	
FV	0.10	0.47	0.42	
Options	*7T*	*7F*	*7?*	
FV	0.34	0.52	0.14	
Options	*8T*	*8F*	*8?*	
FV	0.53	0.36	0.10	
Options	*9T*	*9F*	*9?*	
FV	0.14	0.75	0.11	
Options	*10T*	*10F*	*10?*	
FV	0.30	0.37	0.33	

The actual responses indicated that for items #6, #7, #8, and #10, two or even all three options had FVs within our target range (0.3-0.7). But clearly, this did not mean that two or three options could have been correct. In order to gain further information on which to base a possible revision of the "correct" answers we looked at the Discrimination indices for these items.

The analysis of these responses involved three steps: firstly, we returned to the test and re-checked the items to see whether we continued to find the predicted responses "correct". Items #6 and #9 presented no difficulties as the majority of candidates had chosen the predicted "correct" response. We then looked again at the Actual responses of the other items and rejected some because they could not possibly be "correct". This happened in the case of items #5, #7 and #10, where the predicted key responses were maintained. The only remaining doubts centered on item #8.

During the writing of the test, the response to this item was considered to be within Paragraph D. On re-reading it is clear that "never" is not part of the proposition. The phrase "muy posiblemente, se habrían acortado" (Para. C) conditions the statement to such an extent that the "correct" response must be "False". Consequently, this is reflected in the adjusted "correct" responses.

Figure 13—3 Sub-test es01.2 Items #5-10 Facility values of key responses: a better set of figures than could be hoped for

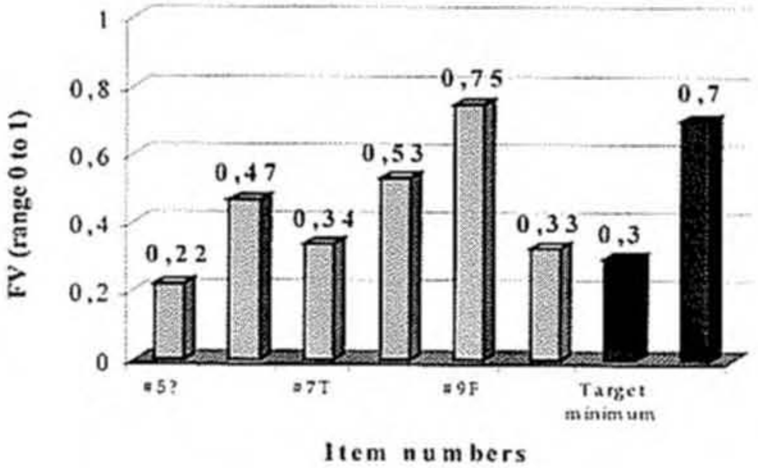
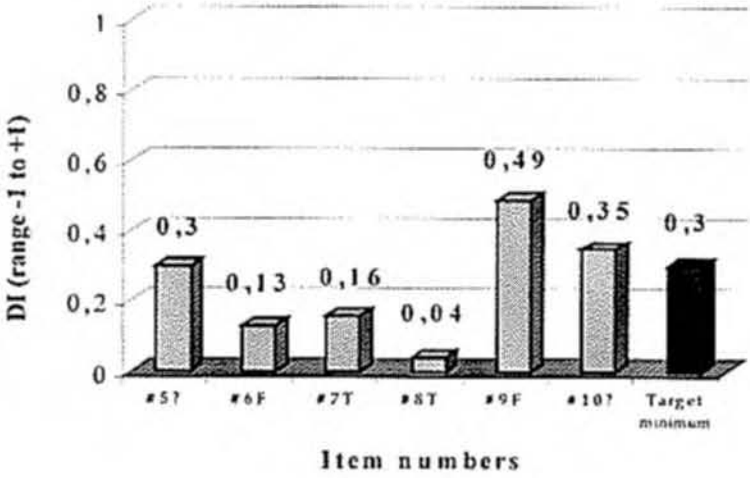


Figure 13—4 Sub-test es01.2 Items 5-10 Discrimination indices of key responses



13.2.2.2 *Discrimination indices*

Half of the options used in these items produced *negative* DIs. This led us to believe that the stronger candidates had in some way been "tricked" by the items and answered incorrectly, while the less strong candidates answered the same items correctly, perhaps as a result of the random choice factor that multiple choice items entail.

13.2.3 *DISCUSSION*

Table 13—8 Items #5-10 Discrimination indices of 3-option items (shading indicates key responses)

Options	*#5T*	*#5F*	*#5?*	***Target***
DI	(0.30)	0.00	0.30	**>0.29**
Options	*#6T*	*#6F*	*#6?*	
DI	(0.14)	0.13	0.08	
Options	*#7T*	*#7F*	*#7?*	
DI	0.16	(0.24)	0.09	
Options	*#8T*	*#8F*	*#8?*	
DI	0.04	(0.08)	(0.01)	
Options	*#9T*	*#9F*	*#9?*	
DI	(0.34)	0.49	(0.15)	
Options	*#10T*	*#10F*	*#10?*	
DI	(0.21)	(0.15)	0.35	

In our revision of this section of the paper we concluded that item #5 had been marginally successful as the key answer produced an FV of 0.22. This was below our preferred range and indicated the item was quite difficult, but with a minimally acceptable DI (0.30) the item seemed to have functioned correctly. We decided to maintain the "correct" answer predicted at the time of writing. Items #6, #7 and #8 were also left unchanged: the FVs produced (0.47, 0.34, 0.53 respectively) were acceptable, but each of them failed to discriminate significantly, with DI values of 0.13, 0.16, and 0.04 respectively. Item #9 was easier than had been hoped (FV 0.75), but provided an acceptable DI at 0.49. This item was perhaps the most satisfactory within this section. Finally, item #10 was also maintained because both the FV (0.33) and the DI (0.35) were minimally within our targets.

In general, though, we found that these items functioned poorly and did not really provide valuable data with regard to candidate ability.

13.2.4 *FIVE-OPTION MULTIPLE CHOICE QUESTIONS*

Table 13—9 Items 11-20 Target skills (Munby 1978:126-131)

Item	*Target skills*
#11, #12	Understanding relations within the sentence (Skill 28)
#13	Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)
#14, #16	Understanding explicitly stated information (Skill 20)
#15	Recognising indicators in discourse (Skill 35)
#17- 20	Deducing the meaning and use of unfamiliar lexical items, through understanding word formation (Skill 19.1) and context (Skill 19.2)

13.2.4.1 *Facility values*

The FV scores produced by these items were high: 0.61-0.83, and although seven out of ten were within our target range, these were all in the top quarter, indicating that all of the items were easier than we would have liked. The other 3 items were above our target maximum, and so we considered them too easy.

13.2.4.2 *Discrimination indices*

As was to be expected, if the FVs tended to be high, and the items were therefore easier than predicted, the DIs tended to be low, and the items did not, therefore, discriminate as much as had been hoped. In fact, the DIs ranged from 0.34 to 0.51, all of which were within the lower half of the target range. None of these items produced a DI outside of our target.

13.2.4.3 *Distractor evaluation*

In general the distractors offered in these five-option multiple choice items performed poorly. In items of this type all four incorrect options, intended to distract candidates from the correct option, would receive >15 % of the responses. However, while the range of FVs for distractors was 0.00-0.33, only 5 out of 40 distractors attracted 15% or more of responses.

13.2.4.4 *Discussion*

The section of multiple choice questions did serve to discriminate between candidates to a certain extent and, given that the candidates were native speakers of the language being tested, perhaps this was as much as could reasonably be expected. Two items produced actual "correct" responses that called into question those predicted initially. After considering the options given, we altered the "correct" responses for items 11 and 13, and changed candidates' scores accordingly.

13.3 Descriptive statistics

Prior to our analysis of the results, we produced three parallel scores for each candidate, based on the three versions of the "correct" responses:

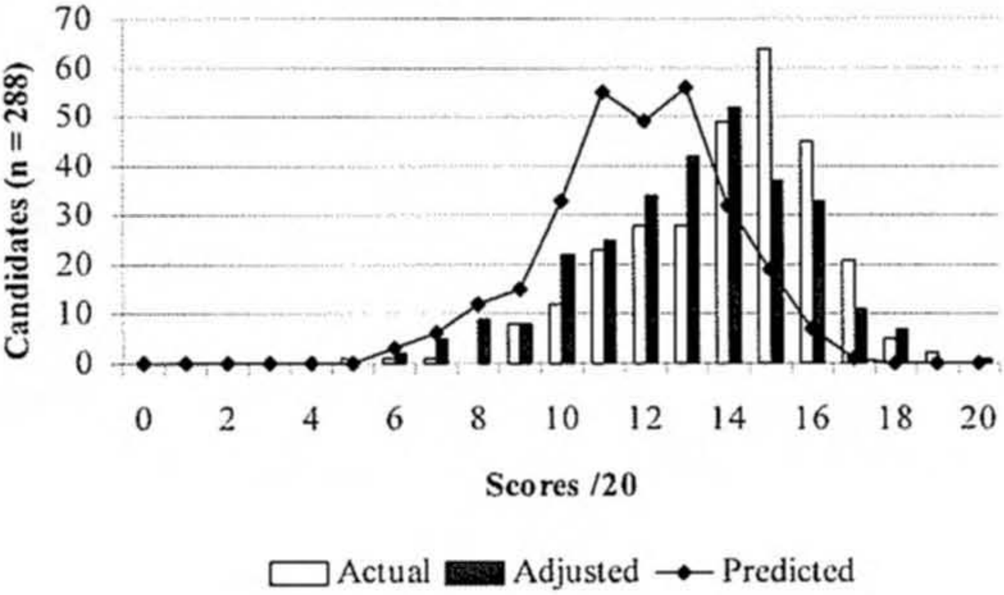
- The key responses, i.e. the answers in the markscheme produced when the test was written
- The actual "correct" responses, i.e. those produced by the majority of the candidates
- The adjusted "correct" responses, i.e. the definitive list produced by authors after reviewing the other two sets of scores in the light of the FVs and DIs.
- Three sets of descriptive statistics were calculated and the frequency distributions plotted for each.

Table 13—10 Descriptive statistics 20 items: actual vs. key vs. adjusted responses

	Key	*Actual*	*Adjusted*
Mean	11.82	13.94	13.14
Median	12.00	14.00	13.00
Mode	13.00	15.00	14.00
Standard error	0.12	0.13	0.15
Standard deviation	2.09	2.28	2.54
Skewness	(0.39)	(0.70)	(0.39)
Level of confidence : 95%	0.24	0.26	0.29

The predicted scores produced a curve with a slight negative skew, but which was the closest of the three to a normal bell-shaped distribution (Figure 13—5). This was also seen in the final adjusted scores, although the skew was more marked here, indicating a comparatively easy test. The actual scores gave a sharper left-hand skew, which indicated that these would have produced the "easiest" version of the paper.

Figure 13—5 Sub-test es01.2 Frequency disribution of the scores achieved by candidates: actual scores vs. key responses vs. adjusted responses



13.4 Reliability

The level of reliability achieved by this sub-test was unsatisfactory, and this meant that the Standard error of measurement was also off-target. As the test contained 20 items, there is a 68% chance that candidates' true scores would be likely to fall within a range of their raw scores ±1.62. The margin of error is clearly too high.

Table 13—11 Split-half reliability coefficients

	Formulae		*Average*
Criterion	*Spearman-Brown*	*Guttman*	
FV	0.529	0.632	0.581
DI	0.553	0.654	0.603
Average	0.541	0.643	
Average overall	0.592		

Table 13— 13 Standard error of measurement (Target <1)

S_e	1.62

The more promising result in terms of reliability comes from applying the prophecy formula. By adding 6 homogeneous items, that is six items of a similar type and with similar FV and DI scores to those in the current version, we would be able to achieve a reliability coefficient of 0.947. This in its turn, would reduce the level of Standard error.

Table 13—12 Improving reliability

Spearman -Brown Prophecy Formula	0.947

13.5 Conclusions

In the analysis of results, conflicts arose between the "correct" responses predicted and those provided by candidates. In most instances, the authors confirmed the predictions after duly analyzing the test papers, descriptive statistics, FVs and DIs. The most conflictive items are #5 to #10, those of the True/False/Don't know type.

This test was based on a text and on items presented in the mother tongue of the candidates and, perhaps as a consequence, it did not provide a wholly reliable means testing them. However, as Table 13—14 shows, thirteen of the twenty items achieved the targets established in terms of FV and DI scores.

Four of these items came from the matching exercise, and these were designed to be easy. We did not expect these items to discriminate between candidates. All of them were targeted on scanning.

Eight of the other nine successful items were MCQs, and their performance was good, although there was little difference between them and we might have hoped for some to be noticeable better discriminators than others in order for us to identify higher and lower levels of discrimination. These items focussed on a range of skills, but it is worth mentioning that four of them were intended to test skill 19:

> *Deducing the meaning and use of unfamiliar lexical items.*

This is regarded by most writers (e.g. Madsen 1983; Weir 1993) as one of the lower level skills. As we are dealing with candidates' Language A, it seems strange to find that a lower level skill should actually provide us with a measure of discrimination.

The remaining item was #10, one of the True/False/Don't know type, and it focussed on "understanding explicitly stated information", a combination that we have already seen in sub-tests es01.1 and in01.2, to work well. This item was both difficult and a good discriminator, and in this it contrasted with all of the others of the same type. The overall performance of this task type led us to conclude that it should be omitted from the revised test specification to be prepared for Cohort B.

Table 13—14 Sub-test es01.2 Item by item summary of performance — shading indicates items meeting both FV and DI targets

Item	*Task type*	*Skill(Munby 1978:126-131)*	*FV*	*DI*
#1	Matching titles to paragraphs	Scanning to locate specifically required information on a whole topic (Skill 46.5)	0.98	0.05
#2			0.97	0.06
#3			0.99	0.08
#4			0.85	0.18
#5	True/false/don´t know	Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)	0.22	0.30
#6			0.47	0.13
#7		Understanding relations between parts of text through the grammatical cohesion device of reference (8 Skill 32.1)	0.34	0.16
#8		Understanding explicitly stated information (Skill 20)	0.53	0.04
#9			0.75	0.49
#10	5-option MCQ	Understanding relations within the sentence (Skill 28)	0.33	0.35
#11			0.62	0.51
#12			0.83	0.34
#13		Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)	0.62	0.44
#14		Understanding explicitly stated information (6 Skill 20)	0.67	0.35
#15		Understanding relations between parts of a text through grammatical cohesion devices of reference (anaphoric and cataphoric) (8 Skill 33.1)	0.83	0.35

Item	*Task type*	*Skill(Munby 1978:126-131)*	*FV*	*DI*
#16		Understanding explicitly stated information (6 Skill 20)	0.80	0.38
#17 #18 #19 #20		Deducing the meaning and use of unfamiliar lexical items, through understanding word formation (6 Skill 19.1) and context (6 Skill 19.2)	0.61 0.66 0.60 0.63	0.34 0.38 0.37 0.45

14 SUB-TEST ES02

14 SUB-TEST ES02

14.1 TEXT
14.2 TASK TYPES
14.2.1 *Matching exercise*
14.2.1.1 Facility values
14.2.1.2 Discrimination indices
14.2.1.3 Distractor evaluation
Item #3
Item #4
Item #5
14.2.1.4 Discussion
14.2.2 *True/False/Don't know*
14.2.2.1 Facility values
14.2.2.2 Discrimination indices
14.2.3 *Five-option multiple choice questions*
14.2.3.1 Facility values
14.2.3.2 Discrimination indices
14.2.3.3 Discussion
Item #19
Item #17
Item #14
Item #18
Item #12
14.3 DESCRIPTIVE STATISTICS
14.4 RELIABILITY
14.5 CONCLUSIONS

THIS TEXT was chosen as the basis for a paper intended to parallel es01.2. The decision to produce two parallel papers for use in the same session was taken at the request of the Examination board for reasons of security and in view of the predicted number of candidates. This entailed the introduction of a measure of Z-VALUES in order to calculate the equivalent scores on the two papers, in addition to the use of the descriptive statistics. By this means, we were able to establish the equivalence of the two papers, and accordingly scale candidates' results so that no one could be said to have taken an "easier" or "more difficult" examination. The results of this procedure are described in detail below.

The text for es02 was chosen and the items written following the same specification as that used for es01.1. The text was chosen independently, but with knowledge of the text that formed the basis of es01.1 and es01.2, and a different author wrote the items. All of the initial work was carried out independently, and the authors then revised each other's work with the aim of bringing the two papers into line as far as possible. Finally, other members of the Examination board studied the papers in order to consider their CONSTRUCT VALIDITY and PARALLEL TEST RELIABILITY. This sub-test was not trialled before its full administration.

The sub-test was administered to 161 (36%) of the 449 individuals making up the first batch of candidates hoping to gain entry to the degree programme. They took the sub-test in July 1993, in the Faculty of Arts building of the University of Granada. The remainder of the candidates in this batch took sub-test es01.2. Sub-tests English, French and German, based on the same specification, were also administered in this session

14.1 Text

The test specification stated that the text would be journalistic in nature, recently published, dealing with a topic of current affairs, and of approximately 600 words. The text in question, which appears in Table 14—1, was taken directly from *El Pais* (22/6/93:8). It was not edited in any way, and the version presented to candidates was a photocopy of the original onto which capital letters had been pasted to identify the different paragraphs. The total number of words exceeded the specification: it was just over 700 words long, more than 200 words longer than sub-test es01.2.

Table 14—1 "Urge la tolerancia" *El País* (22/6/93. EDUCACIÓN:8)

LA ADQUISICIÓN DE VALORES

Unicef intensifica su campaña para educar a los jóvenes como "ciudadanos del mundo"

ANA FERNÁNDEZ

A El Fondo de las Naciones Unidas para la Infancia (Unicef) está llevando a cabo una campaña internacional dirigida a concienciar e integrar en los sistemas educativos una nueva opción pedagógica que es la educación para el desarrollo, y cuyo objetivo fundamental es sensibilizar al ser humano a fin de que colabore en la solución de los problemas y diferencias, tanto económicas como culturales o étnicas, que afectan a nuestra sociedad, ofreciendo así una alternativa frente a la violencia que a menudo se utiliza para dirimir tales diferencias.

B El responsable del Programa de Educación para el Desarrollo de la oficina europea de la Unicef, Andrés Guerrero Feliú, opina que, aunque la educación para el desarrollo no da una respuesta a todos los problemas que surgen en este mundo interrelacionado, "intenta, al menos, crear en los jóvenes las actitudes que les van a permitir comprender los complejos sistemas de relaciones que dominan nuestra sociedad y de cómo éstos influyen en uno mismo en su relación con el entorno".

C La educación para el desarrollo, una de las principales preocupaciones de fondo desde los años 70, está propulsada principalmente, desde los propios comités nacionales que la Unicef tiene en la casi totalidad de los países industrializados que son los que más elementos de juicio tienen a la hora de elaborar material informativo y didáctico a cada país.

D Si en un principio esta inquietud nació para informar y sensibilizar a los jóvenes del Norte sobre los problemas de los países en desarrollo, el nuevo enfoque de la organización está centrado en la promoción de una "ciudadanía global".

E La solidaridad, la paz, la tolerancia, la justicia social, la conciencia de los problemas del medio ambiente son los valores que promueve la educación para el desarrollo. Asimismo, intenta ofrecer a los jóvenes los conocimientos necesarios que les permitan poner en práctica dichos valores, según Guerrero.

F "La guerra es un producto de la mente del ser humano. Si somos capaces de generar violencia, también lo somos de crear armonia. Frente a nuestra capacidad de destrucción debemos explotar la capacidad de construcción".

G Opina que la crisis en la antigua Yugoslavia refleja "el fracaso de la no práctica de la educación para el desarrollo". Esto nos enseña que los problemas "se resuelven sobre todo a nivel local en su interacción con el medio donde se producen".

Ciudadanía global

H Los tejidos de relaciones humanas, que en el mundo de hoy se multiplican aceleradamente gracias al avance de los sistemas de comunicación, hacen que el planeta se enfrente a un complejo proceso de interconexiones que generan situaciones y problemas de naturaleza global.

I Por ello, la Unicef considera imprescindible educar a la juventud y fomentar una visión global, mediante el estudio del planeta y de su gente, de los distinto estilos de vida y de la forma en que se aprehende la realidad mundial.

J En este sentido, los conceptos de interdependencia - aprender a valorar las interconexiones y las consecuencias que muchas acciones efectuadas a nivel local acarrean a nivel mundial-, la exploración de otros modos de vida y puntos de vista, eliminando los estereotipos que son siempre juicios de valor muy simplistas o la justicia social, mediante la adquisición de conocimientos en materia de derechos humanos y su aplicación a la vida diaria, son elementos que forman parte del aprendizaje de la ciudadanía global.

K Asimismo, Guerrero destaca otros conceptos inherentes a este nuevo tipo de ser humano como son la enseñanza a la juventud del origen y la solución de conflictos, para lo que necesita comprender las diferentes fuentes y las causas de los mismos. Debe aprender a luchar por la paz y entender que las medidas que se toman hoy afectarán a las generaciones futuras.
L "Una de las dimensiones de la globalidad es el tiempo", señala; "cada acontecimiento hunde sus raíces en algo que ya se produjo y a su vez influirá en lo que se va a producir. Una preparación para el futuro implica comprender que cada uno de nosotros tiene algo que decir ante las fuerzas que provocan los cambios y una participación activa en el mismo proceso".

No statistical comparison between this text and that used for sub-test es01 was possible as we were not able to use Flesch or Flesch-Kincaid text difficulty software, but neither the authors nor the board members who studied the papers thought there was a significant difference in terms of difficulty. The question of the difference in the number of words was not considered important due to the nature of the texts, the length of time allowed for the test, and the fact that subsequent use of the results would be based on Z-VALUES and not the raw scores.

14.2 Task types

14.2.1 *MATCHING EXERCISE*

The internal construction of this text permitted five matching items, as opposed to four used in sub-test es01.2. The author felt this was justified because the text comprises 12 paragraphs (identified by the letters A to L) whereas es01.2 was only five paragraphs long. The results of trialling es01.1 had led to the inclusion of one distractor paragraph in the former sub-test meaning that only four test items could be used. Consequently, it was considered acceptable to use a text that would give seven distractor paragraphs.

However the extra seven distractors did increase the range of choice, and therefore of possible error, for candidates. In the analysis that follows we will see how this did not actually induce errors, and might even be considered to have had a positive effect on the Facility values. It did not influence the Discrimination indices, however.

The target skill for these items was the same as that for the equivalent items on sub-test es01.2 (Table 14—2).

Table 14—2 Items #1-5 Target skills (Munby 1978:126-131)

Item	*Target skills*
#1-5	Scanning to locate specifically required information on a whole topic (Skill 46.5)

14.2.1.1 *Facility values*

As in the trial and full versions of es01, we aimed for all candidates to find these items relatively easy. In fact, all candidates answered item #2 correctly. However, items #3 and #5 produced FVs of 0.73 and 0.76 respectively, which while still high, do indicate that the items were not all as easy as we had intended (Figure 14—1).

14.2.1.2 *Discrimination indices*

The DI scores indicated that, despite the longer text and the increased number of distractors, these items failed to discriminate significantly between candidates. Three items discriminated to a very limited extent, and one of these produced an unusual set of responses, which we will look at in more detail below (Figure 14—2).

14.2.1.3 *Distractor evaluation*

Given that two of the five items produced FV scores below our target, we looked at the performance of the 12 options, in order to establish whether any pattern existed. Having fixed our target at FV ≥0.8 we would expect the remaining 20% of responses to divide equally among the options. In a perfect test, this would mean that three candidates would choose each of these incorrect options.

ITEM #3

Table 14—3 Item #3 Question stem with the key response in bold italics

3.	"La globalidad de los problemas."	*Para H*

If we look at the set of FV scores produced by all of the options for item #3, which appear in Table 14—4, we can see that option #3I attracted 18 responses, and that two options — #3E and #3K — gained none at all. However, over the remainder there was an even spread. We can probably, therefore, say that this item was

Figure 14—1 Sub-test es02 Items # 1-5 Facility values

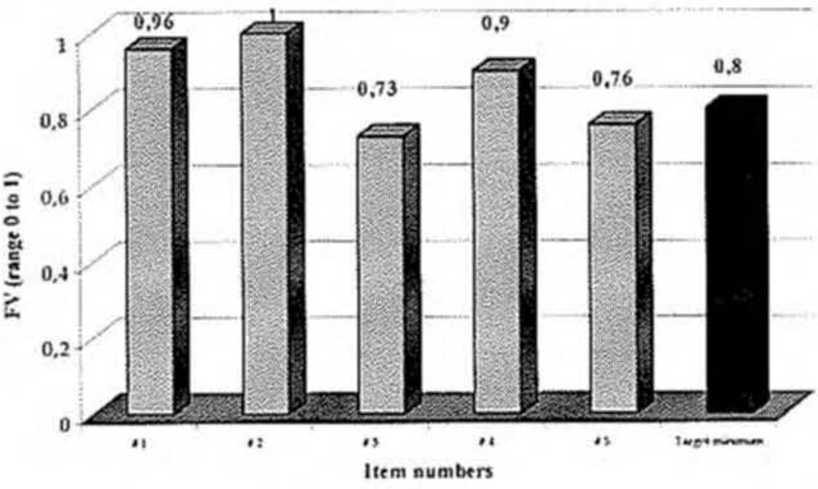
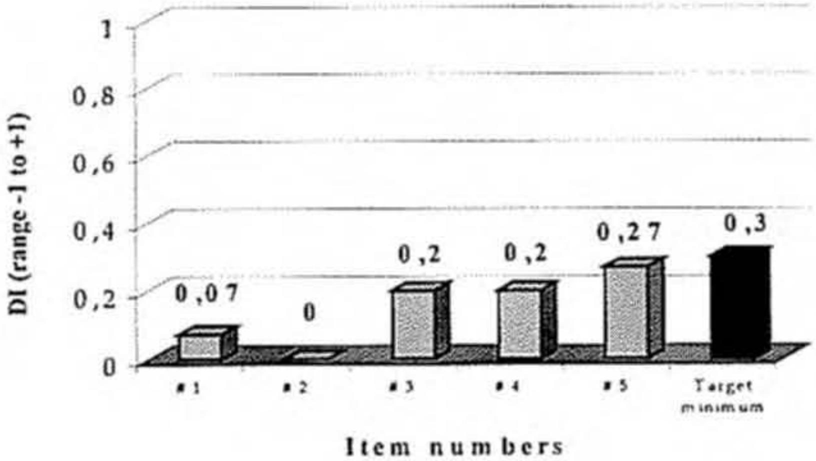


Figure 14—2 sub-test es02 Items #1-5 Discrimination indices



reasonably well constructed. The clues that would lead to a correct answer appear in the second of the two clauses that make up the one-sentence paragraph:

> *...hacen que el planeta se enfrente a un complejo proceso de interconexiones que generan situaciones y problemas de naturaleza global. (Paragraph H)*

These are essentially lexical clues found in the words *problemas* and *global* both of which are repeated, fully and partially, in the question stem.

Table 14—4 Item #3 Distractor performance Facility values

Item	*#3A*	*#3B*	*#3C*	*#3D*	*#3E*	*#3F*
Actual responses	1	6	2	7	0	2
FV	0.01	0.04	0.01	0.04	0.00	0.01
Total responses	163					
Item	*#3G*	*#3H*	*#3I*	*#3J*	*#3K*	*#3L*
Actual responses	3	117	18	3	0	4
FV	0.02	0.73	0.11	0.02	0.00	0.02
Total responses						

ITEM #4

Item #4 produced an FV score of 0.9, and a DI of 0.2. Data on the distractors show that incorrect responses were spread among the options, with three attracting zero responses, and the remainder ranging from 1 to 4. Within our limited expectations of these items, we can say that #4 performed moderately well.

ITEM #5

When we look at the performance of the options that correspond to item #5, we find that only three options attracted any responses at all, and the strongest options were the key response — #5K — and the paragraph immediately following.

In this case, if we look more closely at the item stem we can learn whether the item was satisfactory. Paragraph I, chosen by four candidates, is linked in content, but no more. Paragraph L, is a quotation which supports Paragraph K, and it includes the word

Table 14—5 Item #5 Distractor performance Facility values

Item	*#5A*	*#5B*	*#5C*	*#5D*	*#5E*	*#5F*	*#5G*	*#5H*	*#5I*	*#5J*	*#5K*	*#5L*
Actual Responses	0	0	0	0	0	0	0	0	4	0	122	36
FV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.76	0.22

futuro which the question echoes. However, in Paragraph K, the very same words used in the item stem appear at the end of the paragraph, which makes it difficult to understand why 36 individuals should make this mistake. Were they, perhaps, being "too clever"?

Table 14—6 Item #5 Question stem with the key response in bold italics

5.	"Las generaciones futuras."	*Para K*

Candidates do sometimes assume that if the answer to an item is obvious then this is a trick. If we look at the DI histogram for these items (Figure 14—2), we can see that item #5 was the best discriminator of all (DI 0.27) so it would appear that, even if the distractors failed to operate successfully, the item did serve some purpose. Was this purpose a function of the candidates' linguistic ability? We conclude that it probably was not, and that the discrimination between candidates was more likely to be a factor of test-taking experience or, even, of general intelligence.

14.2.1.4 *Discussion*

Our analysis of these five items teaches one important lesson for the writing of future tests, in that it demonstrates the difficulties that can be encountered in the use of this particular task type. The number of distractors is an important factor, and the exam preparation that candidates may have had can influence the way they respond to the items. The test writer is never likely to be able to cater for this factor when dealing with an unknown population, unlike the classroom teacher. Here we have a clear example of the type of extraneous factor that can influence construct validity and test reliability.

From the point of view of the item writer, a positive point is that apparent value to be gained from using options that repeat fully or partially, phrase in the text.

14.2.2 *True/False/Don't know*

This section contained only five items, as compared to the six items in sub-test es01.2, in order to compensate for the extra item in the previous section. This was in line with the decision that the test writers working on all four languages had taken.

Table 14—7 Items #6-10 Target skills (Munby 1978:126-131)

Item	*Target skills*
#6	Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)
#7-10	Understanding explicitly stated information (Skill 20)

The target skills on which these items focussed were also used in es01.2, but none of the items targeted on

> *Understanding relations between parts of text through the grammatical cohesion device of reference (Munby 1978:126-131 Skill 32.1)*

which was the object of one of the items in the parallel test.

14.2.2.1 *Facility values*

Table 14—8 Items #6-10 Facility values for all options with the key responses shaded

Options	*#6T*	*#6F*	*#6?*
FV	0.41	0.57	0.01
Options	*#7T*	*#7F*	*#7?*
FV	0.35	0.56	0.09
Options	*#8T*	*#8F*	*#8?*
FV	0.12	0.45	0.43
Options	*#9T*	*#9F*	*#9?*
FV	0.93	0.06	0.01
Options	*#10T*	*#10F*	*#10?*
FV	0.25	0.29	0.47

Four of the five items in this section produced FV scores that fell within our target range. If we look at candidate choice concerning the key responses, we see that the FV scores for items #6, #7 and #10 were in the middle of our target range. Item #9 produced a very high FV (0.93); item #8 produced two options with very similar values (0.45 and 0.43).

14.2.2.2 *Discrimination indices*

Similarly, four out of the five items were comfortably above our target score. Moreover, the details of the DI scores for all options provided us with material of interest, as shown in Table 14—8.

Table 14—8 Items #6-10 Discrimination indices for all options with the key responses shaded

Options	*#6T*	*#6F*	*#6?*
DI	0.57	0.54	0.00
Options	*#7T*	*#7F*	*#7?*
DI	(0.36)	0.43	(0.07)
Options	*#8T*	*#8F*	*#8?*
DI	0.50	(0.32)	0.50
Options	*#9T*	*#9F*	*#9?*
DI	0.05	0.00	(0.05)
Options	*#10T*	*#10F*	*#10?*
DI	0.41	(0.20)	0.41

By comparing the FVs and DIs for the options on item #8 we found that the predicted "correct" answer, "?" "there is no evidence in the passage to support this" gave a sound DI, whereas the slightly more popular option, "False", gave a negative DI of (0.32). This meant that generally the weaker candidates had chosen "False".

14.2.3 *FIVE-OPTION MULTIPLE CHOICE QUESTIONS*

Overall, the results derived from this set of items were promising. With the exception of item #12, which was rejected after analysis of candidate responses as defective, the other items performed reasonably well.

Table 14—9 Items #11-20 Target skills (Munby 1978:126-131)

Item	*Target skills*
#11, #14, #16, #17, #19	Understanding explicitly stated information (Skill 20)
#12, 18	Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)
#13	Understanding the communicative value (function) of sentences and utterances without explicit indicators (Skill 26.2)
#15	Understanding relations between parts of text through grammatical cohesion devices of reference (anaphoric and cataphoric (Skill 33.1)
#20	Interpreting text by going outside it, using exophoric reference "reading between the lines" (Skill 34.2)

14.2.3.1 *Facility values*

The range of Facility values achieved by these items was reasonably satisfactory, as Figure 14—3 shows. Four items fell within our target range; three were above it, and two below.

14.2.3.2 *Discrimination indices*

The DIs ranged from 0.05 to 0.70. Four items surpassed our target, while five equalled, or were below the level set (Figure 14—4). Items #14, #17, #18 and #19 all produced acceptable FVs too, and three of them focussed on the same skill.

14.2.3.3 *Discussion*

In this section we will discuss most of the MCQ items and go into some detail in our attempt to explain their success or failure. We will begin with item #19, apparently the most successful on the test paper with a low FV (0.35), and a satisfactory DI (0.38) score. This means that the item is both difficult, and discriminates well. Then we will look at the other items that focussed on the same skill — *understanding explicitly stated information* — not all of which were equally successful.

Figure 14—3 Sub-test es02 Items #11-20 Facility values

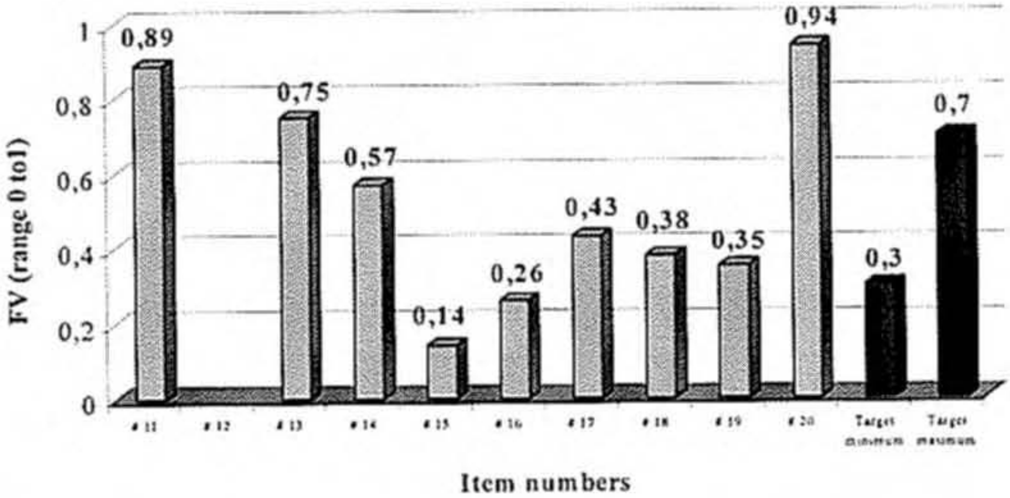
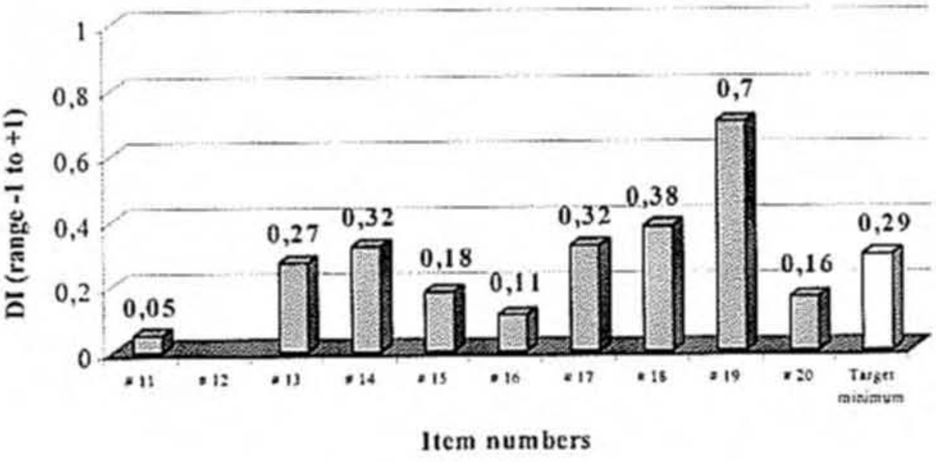


Figure 14—4 Sub-test es02 Items #11-20 Discrimination indices



ITEM #19

Table 14—10 Item # 19 Question stem and all options (a ✓ indicates the key response)

La adquisición de conocimientos en materia de derechos humanos contribuye a ...

A la eliminación de estereotipos. ____
B su aplicación a la vida diaria. ____
C el avance de la justicia social. ✓
D la valoración de las interconexiones. ____
E la exploración de puntos de vista. ____

This item is based on Paragraph J of the text, and has certain characteristics that, at first sight would almost seem to mediate against its success. In the first instance, the head noun phrase of the stem appears in its entirety in the text. Furthermore, all of the options are either quotations or close reflections of phrases, which also appear in the same paragraph. A third factor to bear in mind is that none of the items are mutually exclusive: i.e. at first sight, one or more of the options could be correct.

On closer analysis of the options and the text, we find that the secret of the item appears to lie in the complexity of the sentence or paragraph structure of the text. This may well be a good example of the item writing opportunities that Spanish offers through its propensity for lengthy, complex sentences involving many layers of subordination. Similar opportunities are unlikely to appear as often in an English text.

However, as Table 14—11 shows us, the number of responses attracted by each of the distractors in this item was uneven. Ideally, with an FV of 0.35 for the key response — option #19C — we would expect each of the alternatives to score around FV 0.16. In fact, two items score slightly less than this, but the other two were unbalanced. Distractor #19A is clearly too strong, attracting 52 candidates and achieving an FV of 0.32.

Table 14—11 Item #19 Facility values for all options with the key response shaded

Item	*#19A*	*#19B*	*#19C*	*#19D*	*#19E*
Actual responses	52	8	56	23	21
FV	0.32	0.05	0.35	0.14	0.13

ITEM #17

To continue to focus on the items targeted at understanding explicitly stated information, we now turn to #17, reproduced in the table, which is based on Paragraph I of the text.

Table 14—12 Item #17 Question stem and all options (a ✓ indicates the key response)

La Unicef considera imprescindible fomentar ...

A el estudio de los distintos modos de vida. ______

B una visión global de las formas de aprehender la realidad mundial. ✓

C la educación de la juventud para aprender de la realidad mundial. ______

D una visión global del planeta y su gente. ______

E la educación de la visión de la juventud ______

Many aspects of the item are similar to #19: the stem and the options are lifted directly from, or reflect very closely, the phrasing of the text. The paragraph itself, while shorter than Paragraph J similarly contains one sentence only.

Table 14—13 Item #17 Facility values for all options with the key response shaded

Item	*#17A*	*#17B*	*#17C*	*#17D*	*#17E*
Actual responses	6	70	32	35	18
FV	0.04	0.43	0.20	0.22	0.11

The performance of the distractors is also similar to that described in relation to #19, although slightly better. A balanced set of options would have attracted about 22 responses each, giving FV scores of 0.14. In fact, two options attracted more responses than this, but neither approached that of the key response. The remaining two were less popular, with #17A having an almost negligible impact on the item.

ITEM #14

The third successful item, which focussed on Skill 22, was #14. This item is based around Paragraph A of the text.

Table 14—14 Item #14 Question stem and all options with a ✓ to indicate the key response

La educación para el desarrollo es un proyecto ...

A de países industrializados. ______

B de nueva opción pedagógica. ✓ ______

C de las Naciones Unidas. ______

D centrado en la infancia. ______

E en los países del Norte. ______

The nature of the stem and of the options is different to the two items discussed above. This may well account for the fact that the FV score was 0.57, notably higher than the others were. The DI value, however, remains in the same band.

If we look at the options in turn we can see that #14A and #14E cancel each other out, in that the two alternatives could in fact be considered synonyms. Neither expression appears directly or indirectly in the paragraph, but both appear later in the text, in separate contexts. It would seem that the writer's intention was more that candidates should choose one or other of these responses if answering in haste, or with time for only a superficial reading of the text.

Table 14—15 Item #14 Facility values for all options with the key response shaded

Item	*#14A*	*#14B*	*#14C*	*#14D*	*#14E*
Actual responses	14	91	50	6	2
FV	0,09	0,57	0,31	0,04	0,01

Options #14C and #14D are similar, but clearly better alternatives. Both repeat words that appear in the paragraph, and could be chosen as the correct response by readers who spend insufficient time on the item. The FV scores they generate support the improved quality of these options.

ITEM #18

The next item we will discuss was based on the skill of

> *understanding information in the text, not explicitly stated, through making inferences (Munby 1978:126-131 Skill 20)*

and the response to the item is couched in Paragraph K. The question stem and options appear in the table.

Table 14—16 Item #18 Question stem and all options (a ✓ indicates the key response)

Los conceptos inherentes al nuevo tipo de ser humano son los conceptos _________ ser humano.

A	... coaligados al ...	
B	... colaterales del ...	
C	... circunstanciales del ...	
D	... adicionales al ...	
E	... esenciales del ...	✓

On studying the candidates' responses we conclude that, essentially, the correct response to this item is to be found in adequately distinguishing between the lexical items included in the options as much as in a correct reading of the paragraph. While the knowledge of vocabulary is obviously a part of all individuals' reading skills, in this case it is not a skill that is being tested by reading in context. However, having said that, the item was of average difficulty, appears to have had no serious internal defects, and discriminated between candidates.

The FV scores of the incorrect options show that options #18C and #18D performed reasonably well, but that #18A was too strong a distractor, at the expense of #18B. Why is this the case? There were no echoes or repetitions in the text that might have brought this about, and we concluded that candidates may have opted for #18A as much because it is the least frequent of the words used, as for any other apparent reason.

Table 14—17 Item #18 Facility values for all options with the key response shaded

Item	*#18A*	*#18B*	*#18C*	*#18D*	*#18E*
Actual responses	50	1	27	18	61
FV	0,31	0,01	0,17	0,11	0,38

ITEM #12

As we will now explain, this item was defective and was discarded from our final analysis of the sub-test. However, as this analysis was not carried out before producing the scores used to admit, or not candidates, it did form part of the scores used for this purpose.

The actual "correct" responses to item #12 indicated that the key option #12D was incorrect. In marking the scripts we found that equal numbers of candidates adjudged both #12A and #12D correct.

When we reviewed Paragraph B, we decided that both options were correct, so this item was eliminated from the final scores and the adjusted answer list was based on totals out of 19, scaled to 20.

Performance of items #14, #17, #18, and #19 - three of which aimed to test the same target skill - indicated that the item type and target skill were appropriate for inclusion in further tests.

14.3 Descriptive statist ics

The frequency distribution for the final set of scores on this sub-test was"normal" (Figure 14—5). The three central measures were all very similar, and the degree of skewness was small. This indicates that the test was, if anything, slightly easy for the majority of the candidates. The negative tail seen in the figure is notable for the sharp fall between the number of candidates scoring 10 out of 20, and 9 out of 20. The total number of candidates who achieved scores below ten did not reach 20, out of 161 who took the paper.

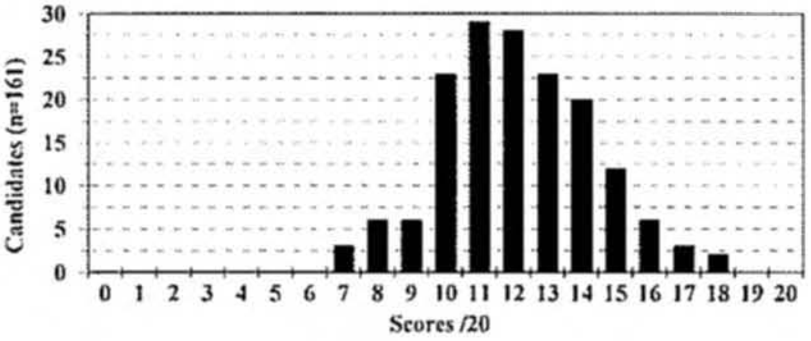
Table 14—18 Descriptive statistics based on scores out of 20: the comparison shows only the slightest difference between the three

	Key	*Actual*	*Adjusted*
Mean	12.04	12.91	12.56
Median	12.00	13.00	12.63
Mode	11.00	12.00	13.68
Standard error	0.19	0.19	0.20
Standard deviation	2.42	2.45	2.54
Skewness	(0.56)	(1.03)	(0.62)
Level of confidence: 95%	0.37	0.38	0.39

14.4 Reliability

Despite the relative "normality" of the frequency distribution (Figure 14—5), the reliability coefficient was most disappointing, with the overall average far lower than we had hoped for (Table 14—19). In order to improve on this aspect of the test we would have to use half as many items again, as were in the current

Figure 14—5 Sub-test es02 Frequency distribution of raw scores out of 20



version, and this seems rather an unrealistic figure. The text might be expected to provide material for another ten items, but it is unlikely that they would be of the same level in terms of the FV and DI scores

Table 14—19 Calculations of split-half coefficients of reliability

	FV rank-order	*DI rank-order*	*Average*
Spearman-Brown			
rxx1	0.389	0.356	0.372
2rhh1	0.482	0.434	
rhh1	0.241	0.217	
1+rhh1	1.241	1.217	
Guttman			
rxx1	0.544	0.516	0.530
s2h1	2.187	2.139	
s2h2	1.809	1.935	
s2x	5.492	5.492	
Average	0.466	0.436	
Overall Average	0.451		

Table 14—20 Standard error of measurement

S_e	1.74

Table 14—21 Reliability of a lengthened version

Spearman-Brown Prophecy Formula (30 items)	0.903

14.5 Conclusions

Our analysis of Sub-test es02 leaves us with thirteen items that performed within our target range for both FVs and DIs, although five of these were the matching items, and as in paper es01.2, these failed to discriminate between candidates. In the section of multiple choice items, we found one that had to be discounted as candidates demonstrated that two of the five options were equally correct.

As Table 14—2 shows, the matching items achieved high FV, and low DI scores, with the exception of #5 which almost reached our target. All of these were based around the same scanning skill.

In contrast to es01.2, four of the five True/False/Don't know items were successful. Three of these were focussed on the skill of understanding explicitly stated information, and the fourth on making inferences. These results are the most positive for this item and skill combination in all of the tests we have analyzed.

Among the MCQ items which reached our performance targets, three were based on the skill of understanding explicitly stated information, skill 20. It appears that this skill and item combination is also predictably successful. However, we have to ask ourselves whether this is a good combination in its own right, or whether it is a good combination because we have mastered the technique of writing this type of item.

Table 14—22 Sub-test es02 Item by item summary of performance

Item	*Task type*	*Skill (Munby 1978:126-131)*	*FV*	*DI*
#1	Matching titles to paragraphs	Scanning to locate specifically required information on a whole topic (Skill 46.5)	0.96	0.07
#2		Idem	1.00	0.00
#3		Idem	0.73	0.20
#4		Idem	0.90	0.20
#5		Idem	0.76	0.27
#6	True/false/ no evidence	Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)	0.57	0.54
#7		Understanding explicitly stated information (Skill 20)	0.56	0.43
#8		Idem	0.43	0.50
#9		Idem	0.93	0.05
#10		Idem	0.47	0.41
#11	5-option MCQ	Understanding explicitly stated information (Skill 20)	0.89	0.05
#12		Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)		
#13		Understanding the communicative value (function) of sentences and utterances without explicit indicators (Skill 26.2)	0.75	0.27
#14		Understanding explicitly stated information (Skill 20)	0.57	0.32

Item	*Task type*	*Skill (Munby 1978:126-131)*	*FV*	*DI*
#15		Understanding relations between parts of a text through grammatical cohesion devices of reference (anaphoric and cataphoric) (Skill 33.1)	0.14	0.18
#16		Understanding explicitly stated information (Skill 20)	0.26	0.11
#17		Idem	0.43	0.32
#18		Understanding conceptual meaning (Skill 24)	0.38	0.38
#19		Understanding explicitly stated information (Skill 20)	0.35	0.70
#20		Interpreting text by going outside it, using exophoric reference "reading between the lines" (Skill 34.2)	0.94	0.16

15 SUB-TEST IN02

15 SUB-TEST IN02

15.1 TEXT
15.2 TASK TYPES
15.2.1 *Matching exercise*
15.2.1.1 Facility values
15.2.1.2 Discrimination indices
15.2.1.3 Discussion
Item #3
Item 4
Item #5
15.2.2 *True/False/Don't know*
15.2.2.1 Facility values
15.2.2.2 Discrimination indices
15.2.2.3 Discussion
Item #6
Items #7 and #10
Items #8 and #9
15.2.3 *Five-option multiple choice questions*
15.2.3.1 Facility values
15.2.3.2 Discrimination indices
15.2.3.3 Discussion
Item #11
Item #14
Item #13
Item #19
Item #20
Item #17
15.3 DESCRIPTIVE STATISTICS
15.4 RELIABILITY
15.5 CONCLUSIONS

THE SECOND LB English sub-test was written to be administered for the first time to the candidates registered for the entrance test in July 1993. This paper was not trialled, and was to be used in parallel with the revised sub-test in01.1. The latter was only to be used if the number of candidates exceeded the room capacity, but as this did not occur, sub-test in02 paper was administered to all 317 in a single session that took place in the Faculty of Arts building of the University of Granada. These candidates also took either sub-test es01.2 or es02. A few also sat sub-tests in German and/or French during a subsequent session. These were candidates who either wished to compete for places in the Applied Languages Europe (A.L.E.) program for which they are required to achieve entrance standard in both English and French, or who hoped to secure themselves a place in the degree program regardless of the LB.

Although the sub-test had not previously been trialled, members of the Examination board had studied it and considered it suitable. Revisions had been made because of their comments.

Our analysis of the results derived from this test is, at least statistically well founded in that the number of candidates who sat the paper was 317. Such a large sample clearly means that the conclusions we draw and any generalizations we may make have a much more sound foundation than those based on sub-test in01.1 (n = 37).

15.1 Text

The text (*The Guardian* 1993) was unedited and the version presented was a photocopy onto which letters had been pasted to identify the paragraphs. It exceeded the specification at 538 words. The topic was of current affairs, required no specialist knowledge, and was thought unlikely to favor any particular group of candidates.

15.2 Task types

The paper involved items of the same three types that were used in all of the papers written for this session: five matching exercises, five True/False/Don't know items, and ten multiple-choice questions. The target skill being tested was a scanning skill, as in all of these matching activities (Table 15—2).

Table 15—1 "Today's royal bride: subversive modern or anti-feminist reactionary?" (*The Guardian* 1993)

Japan marries self-sacrifice with nostalgia

By Ian Buruma

A MODERN Japanese monarchs have liked to copy the style of the British royal family. The late Emperor Hirohito, in particular, admired the casual, golf-playing dash of Edward VIII. He was perhaps a rather odd role model but, his abdication and other problems notwithstanding, British royalty remained until recently the ne plus ultra of monarchical grandeur.

B But Crown Prince Naruhito and Masako Owada marry against a background of changed attitudes to the British monarchy. Now it is viewed by Japanese royalists as a parable of failure, an example of how things can go terribly wrong. Look, for example, at the difference between the Princess of Wales and Miss Owada, a commoner, whose wedding to Prince Naruhito today will make her crown princess.

C The Princess of Wales, one is told (and not only by Japanese conservatives), is the perfect example of modern selfishness. By refusing to play the game, by insisting on her freedom, she has jeopardised an ancient institution. How noble, in contrast, seems Miss Owada: after refusing for seven years to give up a promising diplomatic career for the gilded prison of the Japanese imperial court, she was prevailed on to put aside "selfish" concerns and marry a man for whom she has not expressed any affection. That it was an act of self-sacrifice is clear from her language. While once she said that she would never give up her freedom, she now promises to make herself "useful to the imperial household". Self-sacrifice, for the country, for a feudal lord, for a business corporation, for a husband, is the greatest traditional virtue in a Confucian society, or indeed in any society where the strong can impose virtues on the weak. It is a virtue which fills some Western reactionaries with nostalgia.

D Such nostalgics like to project the shortcomings of their own societies (the selfish, materialistic West) on to other, faraway places (the disciplined, spiritual East). But Miss Owada's virtue is oddly out of step with trends in Japan too.

E Her marriage comes when, for the first time in Japanese history, women are doing what the Princess of Wales has done: claiming the right to disentangle themselves from intolerable marriages. More Japanese women than men now ask for divorces. An increasing number of women are more interested in careers and economic independence than in traditional family life. Some (known derisively as "yellow cabs" - you can always get a ride) would rather travel abroad for casual sex than enter into subservient relationships with Japanese men. One could call this selfish and sad. One could also say that society has not kept pace with its most intelligent and independent women.

F Although there was never any hint of the yellow cab about her, Miss Owada was such a woman. This may be partly why Prince Naruhito was attracted to her, rather than to some simpering blueblood who played by all the old rules. This is what makes the marriage so interesting: either Miss Owada's sacrifice is a reactionary blow against female independence, or she is a subversive modern in the heart of the nation's still-sacred institution.

Ian Buruma is the author of A Japanese Mirror.

15.2.1 *MATCHING EXERCISE*

Table 15—2 Sub-test in02 Items •1-5 Target skills (Munby 1978:126-131)

Items	Skill
#1-5	Scanning to locate specifically required information on one point, involving a complex search (Skill 46.2)

This text consisted of 6 paragraphs, identified by the letters A to F, and five matching items were prepared, thus ensuring that one paragraph served as a distractor (Table 15—3). This contrasted with in01.1, for which there were seven distractors.

Table 15—3 Items #1-5 Question stems

Scan the text and identify for which paragraph each of these sentences could be the title.

1	A distant admirer.	A
2	Contrasting attitudes.	C
3	Out of step with society?	D
4	The enigma of Miss Owada's decision.	F
5	A view from the present.	E

This task differed from that used in the trial test in that the stem included the use of the word "could", and this was underlined. We intended this to allow candidates to speculate and to choose whichever they thought was the best fit.

15.2.1.1 *Facility values*

The target range for these items was ~0.80, and the FV scores produced were in fact lower than that with the exception of item #1, which scored 0.91. This indicates that most of the items were in fact more difficult than had been intended (Figure 15—1).

15.2.1.2 *Discrimination indices*

Having set a high FV range, we expected these items to discriminate little. As the actual FV scores have been lower than our target, some minimal degree of discrimination might be expected from those items that proved slightly more difficult.

To begin with, we can see from the histograms (Figure 15—1 and 15—2) that item #1 was both easy and failed to discriminate. Item #2 was more difficult, but it also scored a mere 0.09 for discrimination. What is more surprising, however, is that item #3 discriminated negatively. This means that the stronger candidates actually chose the wrong response, while the weaker ones were successful. We will discuss this item in more detail below.

Perhaps the only item that performed as might have been expected, in view of the FV score recorded is item #4. This achieved a minimal level of discrimination among candidates, at DI 0.18.

Within this set of items, #5 scored DI 0.31, which is just above our target minimum. If we combine this with an FV of 0.63, we can see that this item performed well. We will look at the reasons behind this in more depth shortly.

15.2.1.3 *Discussion*

One of the clearest conclusions we can draw from our analysis of these items is that, whatever the individual target skill for which the item is written, the realities of candidates' actual responses

Figure 15—1 Sub-test in02 Items #1-5 Facility values

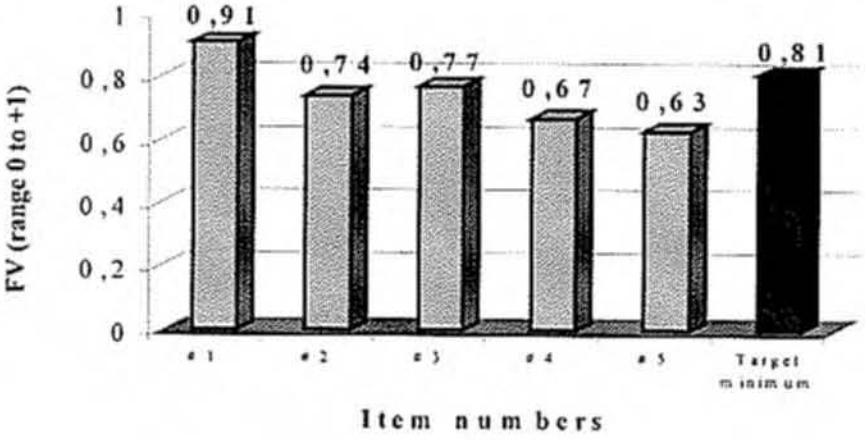
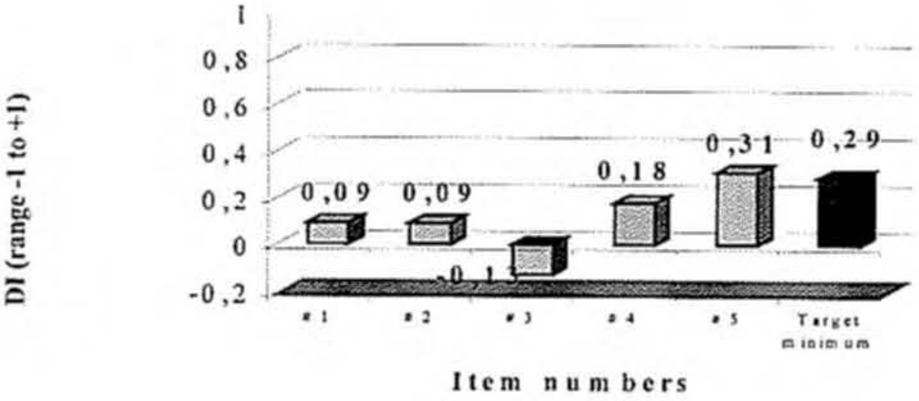


Figure 15—2 Sub-test in02 Items #1-5 Discrimination indices



will produce a much more detailed understanding of the possible range of skills that they may have employed. Reading comprehension is far more complex than the discrete analysis of skills might lead us to suppose.

ITEM #3

As we have mentioned earlier, this item was more difficult than we predicted it to be, and it discriminated negatively. This means that the stronger candidates answered it incorrectly, and weaker candidates gave the correct response.

The item (Table 15—3) is based on candidates being able to link the phrase

> *Out of step with society?*

which appears in the stem, to the phrase

> *...out of step with trends in Japan, too.*

that appears in the text (Table 15—1, Para. D). In order to find out why the better candidates made a mistake with this item we need to look at the responses to all of the options to analyze the effect of the distractors (Table 15—4).

Table 15—4 Sub-test in02 Item #3 Facility values for all options (shading indicates the key response)

Item	#3A	#3B	#3C	#3D	#3E	#3F
Actual responses	2.00	10.00	5.00	242.00	53.00	5.00
FV	0.01	0.03	0.02	0.77	0.17	0.02

The spread of responses across the five incorrect options is by no means even. Option #3E has attracted the attention of 53 candidates, and if we refer back to the text we can see that the entire paragraph does deal with this same topic. There is a phrase in the final sentence that could be said to echo the item stem:

> *...[Japanese] society has not kept pace with its most intelligent and independent women.*

but which is not as close as those that appear in paragraph D

We would suggest that this is an example of an item that better candidates have answered incorrectly precisely because they are better and have been able to complete the exercise more quickly and read the text in more depth. Among the 53 candidates who chose #3E, there will probably be a number who read the text carefully and who decided that the key response, paragraph D, was too obvious. Consequently, they chose E. Less able candidates will have had insufficient time, and will have lacked the capacity to read, and probably re-read paragraph E in order to interpret the item in this way.

ITEM #4

The FV and DI data collected on this item are not in line with our prediction, but in fact make this into one of the better items on the test. The FV score achieved is below our target for the matching exercise, and thus falls into our target range for the other items in the test. However, the DI score remains below our target (Figure 15—2). Why the item performs in this way remains to be seen.

If we look at the FV scores achieved by all of the options we can see that, as in the case of item #3, that the spread among the distractors was uneven (Table 15—5).

Table 15—5 Sub-test in02 Item #4 Facility values for all options (shading indicates the key response)

Item	*#4A*	*#4B*	*#4C*	*#4D*	*#4E*	*#4F*
Actual responses	2.00	0.00	78.00	3.00	20.00	212.00
FV	0.01	0.00	0.25	0.01	0.06	0.67

Alternative #4C was selected by 78 candidates, giving it an FV of 0.25, whereas none of the other distractors reached 0.10. In the view of the item writer, the stem contains the word *enigma* and it seems likely that this is the reason candidates have made an error. Paragraph C deals in much detail with matter of Miss Owada's decision, but in general it does so in affirmations:

> *How noble, in contrast, seems Miss Owada...*
>
> *That it was an act of self-sacrifice is clear from her language.*

In paragraph F, the there is a more open questioning of her motives:

> *This may be partly why Prince Naruhito was attracted to her...*

In this case, the greater capacity of some candidates to read into the details of nuance in the text is what has discriminated between them.

ITEM #5

Like the previous item, #5 has achieved an FV score within our target range for the main body of items. Moreover, the DI score is above our target minimum, making this a valuable item within the overall context of the sub-test (Figures 15—1 and 15—2).

If we consider the construction of this item, we can perhaps learn why it has performed so well. The stem uses the noun *view* as a conscious echo of the verb *viewed* which appears in paragraph B. In the text, *viewed* is also linked to *Look*, which belongs to the same field. The proposition presented in the option draws on the contrast established between paragraphs A and B.

Table 15—6 Sub-test in02 Item #5 Facility values for all options (shading indicates the key response)

Item	*#5A*	*#5B*	*#5C*	*#5D*	*#5E*	*#5F*
Actual responses	6.00	87.00	4.00	4.00	198.00	19
FV	0.02	0.28	0.01	0.01	0.63	0.06

As in the previous two items, the data provided by considering the FV scores of all options (Table 15—6) show us that one of the distractors was far stronger than all of the others. In this case, alternative #5B, the one true distractor in the set, has drawn the attention of 87 candidates, and in fact this can be seen as justified because of the contrast established between this paragraph and the previous one. However, the FV and DI scores attained by the item led us to maintain #5E as the key response, while acknowledging that the item was possibly misleading.

15.2.2 *TRUE/FALSE/DON'T KNO W*

Drawing on the experience gained in the trial tests, this section contains items which combine the task type with skills that have demonstrated themselves to provide us with valuable results (Table 15—7).

Figure 15—3 Sub-test in02 Items #6-10 Facility values for key responses

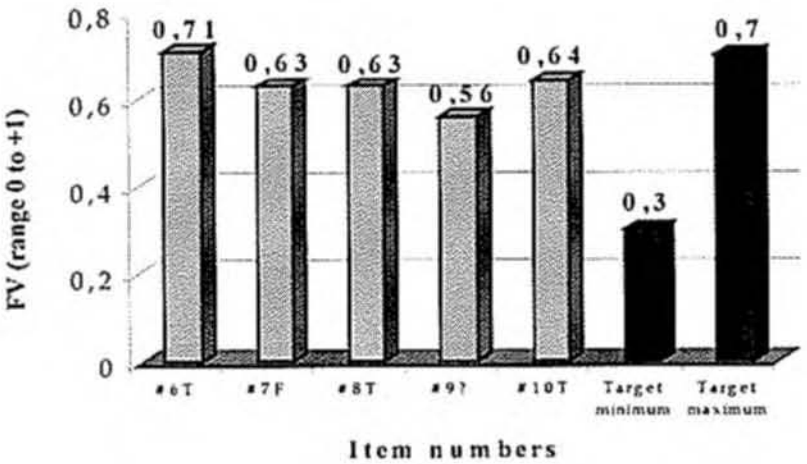


Figure 15—4 Sub-test in02 Items #6-10 Discrimination indices

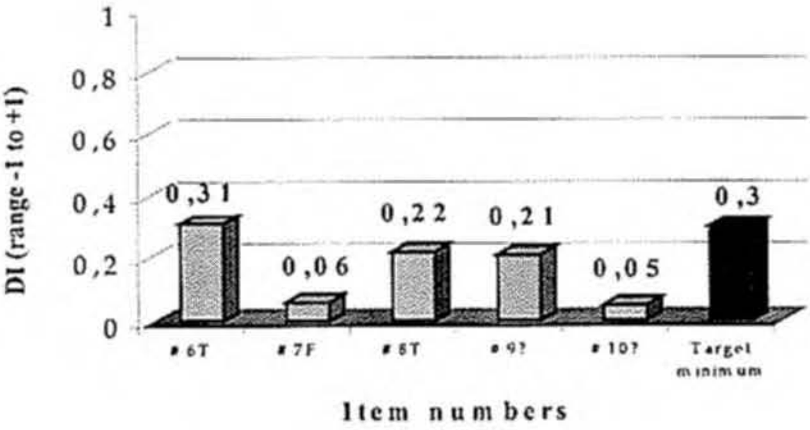


Table 15—7 Sub-test in02 Items 36-10 Target skills (Munby 1978:126-131)

Item	*Skills*
#6	Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)
#7	Understanding relations between parts of text through the grammatical cohesion device of reference (Skill 32.1)
#8-10	Understanding explicitly stated information (Skill 20)

Table 15—8 Sub-test in02 Items #6-10 Question stems with key responses in bold italics

Choose T (True), F (False), or ? ("there is no evidence in the passage to support this").

6. The Japanese see the Princess of Wales as an example of modern social change. ***T***
7. Miss Owada decided to accept her arranged marriage out of a sense of tradition. ***F***
8. Some Western observers would wish to impose Confucian traditions on their own societies. ***T***
9. Prince Naruhito has spoken out against the attitudes displayed by the British Royal Family. ***?***
10. The stability of Japanese family life has been shaken by the so-called "yellow cabs". ***?***

15.2.2.1 *Facility values*

The levels of the FV scores, with the exception of item #6 are all at the higher end of our target range, and #6 is only just above this (Figure 15—3). This indicates that, although there is no variety in the level of difficulty of the items, almost all have performed as hoped.

15.2.2.2 *Discrimination indices*

Only item #6 discriminated well between candidates. Two others gave low positive values, #7 and #10, and #8 and #9 did discriminate to a limited extent. When we analyzed the FVs and DIs we determined not to alter any of the key responses, and judged the performance of item #6 as satisfactory, #8 and #9 acceptable, and #7, and #10 inadequate. In the following sections, we will look at the possible causes of the performance of these items.

15.2.2.3 *Discussion*

ITEM #6

This item almost achieved both of the target parameters established. The FV score of 0.71 was just outside of our range, but the DI value was on target (Figures 15—3 and 15—4). The item has proved relatively easy, but a good means of discriminating between candidates. The question stem, shown in Table 15—8, is intended to lead readers to the first four lines of paragraph B, which clearly agrees with the proposition, and leads readers to the conclusion that this is true.

The performance of this item reinforces our opinion that the combination of this task with skill 22.1 is valuable in item writing.

ITEMS #7 AND #10

In both of these items, we have very low DI values, which make us query the premise on which the items were written. Candidates generally chose the key responses to these items, but among those who chose other options there was no significant distinction between the stronger and the weaker candidates. In response to this, we ask ourselves whether, again, this is evidence of the difficulty implicit in the use of the "Don't know" option.

The key response to item #7 is "False", because the stem offers neither of the suggested reasons for her decision (paragraph F). Similarly, the "Don't know" response is correct for item #10, because at no point in the text is there any information about the influence of the *yellow cabs* on Japanese society at large.

ITEMS #8 AND #9

In the data produced by these items, we found average FVs, of the same strength as those in items #7 and #10, and low DI scores, but scores that did indicate a degree of discrimination.

The response to the first of these items is explicit in the final sentence of paragraph C. The response to #9 can only be "Don't know", as there is no information given about Prince Naruhito's view of the British royals anywhere in the text.

If these two pairs of items had been grouped differently, we might have been able to come to a more useful conclusion. For example, if the two "Don't know" responses had scored the same or similar DI values, we would have been able to deduce something evident about their performance. As things stand, we are left to search for further data.

15.2.3 *FIVE-OPTION MULTIPLE CHOICE QUESTIONS*

Seven of the ten items in this section focus on basic skills of understanding information, whether explicitly stated or implicit in the text (Table 15—9). In addition, one of the remaining three focuses on deducing the meaning of vocabulary, which is also a skill that has demonstrated itself to perform well in these items.

Table 15—9 Sub-test in02 Items #11-20 Target skills (Munby 1978:126-131)

Item	*Skill*
#11	Understanding explicitly stated information (Skill 20)
#12	Understanding information in the text, not explicitly stated, through making inferences; figurative language (Skill 22.1 & 22.2)
#13	Idem
#14	Understanding explicitly stated information (Skill 20)
#15	Understanding information in the text, not explicitly stated, through making inferences; figurative language Skill 22.1 & 22.2)
#16	Idem
#17	Interpreting text by going outside it, using exophoric reference "reading between the lines" (Skill 34.1 & 34.2)
#18	Understanding information in the text, not explicitly stated, through making inferences; figurative language (Skill 22.1 & 22.2)
#19	Deducing the meaning and use of unfamiliar lexical items, through understanding word formation (Skill 19.1) and context (Skill 19.2)
#20	Understanding relations between parts of text through the grammatical cohesion devices (Skill 32)

15.2.3.1 *Facility values*

The FV scores for these items range from 0.12 to 0.87, but six of the set do fall within our target range, and two others are only slightly above the maximum. The six successful items all prove reasonably difficult, and four of them have DI scores that are also above our target (Figure 15—5).

Figure 15—5 Sub-test in02 Items #11-20 Facility values

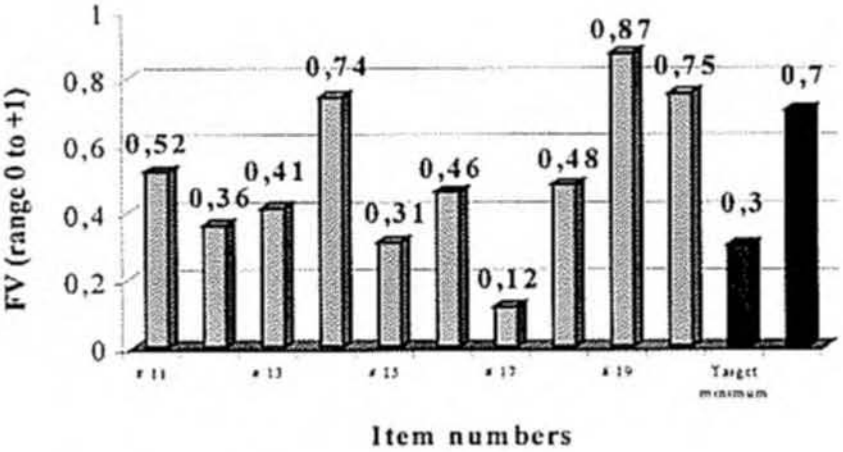
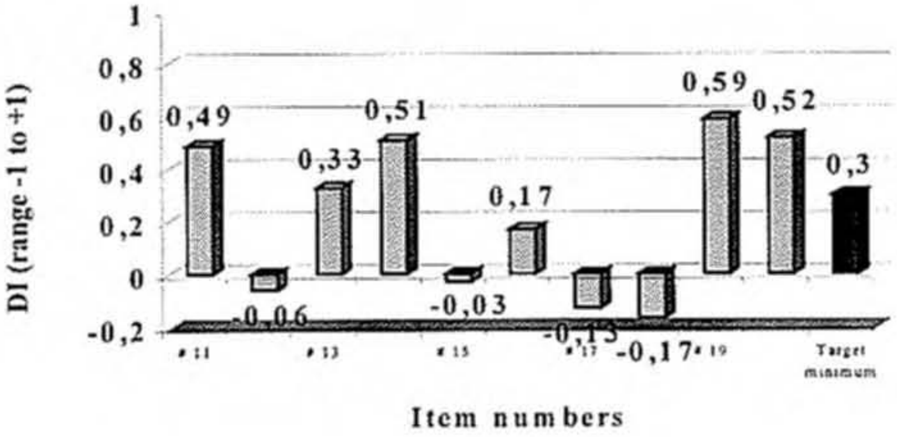


Figure 15—6 Sub-test in02 Items #11-20 Discrimination indices



15.2.3.2 *Discrimination indices*

Five items gave satisfactory DI values, above our target minimum, and the other five were lower than hoped (Figure 15—6). Two of these achieved a level of minimal discrimination, but three produced negative scores. Two of these had also been difficult for candidates, although the third had achieved a reasonably good FV score of 0.48.

15.2.3.3 *Discussion*

In general the MCQs performed reasonably well. By pairing FVs and DIs we find that items #11 and #13 gave the best results, being both reasonably difficult and discriminating well between candidates. Items #14, #19 and #20 produced DIs >0.30, but these were combined with relatively high FVs (range 0.74-0.87) indicating that they probably only served to discriminate at the lower ability levels.

ITEM #11

This item was testing skill 20 (Table 15—9), based on understanding information explicitly stated in the text.

Table 15—10 Sub-test in02 Items #11 Question stem and options (a ✓ indicates the key response)

11. Miss Owada's decision represents...

A a clear change of mind. ✓

B the fulfilment of her dreams. ____

C an acquiescence to family pressure. ____

D her acceptance of destiny. ____

E her desire to be subversive. ____

The answer to this item lies in the second half of paragraph C (Table 15—1), where the writer says

> *While once she said that she would never give up her freedom, she now promises to make herself "useful to the imperial household".*

If we look at the range of options and the numbers of responses each attracted we see that the spread was quite good, with only option #11B failing to gain much support, and this is probably because it is the most obviously incorrect. The option goes completely against the tone of the entire article. The other three options all echo aspects of the text, and partial evidence to support each of them could be found in an incomplete reading of the text. We judge this item to be an extremely successful means of discriminating between candidates.

Table 15—11 Sub-test in02 Item #11 Facility values for all options (shading indicates the key response)

Item	*11A*	*11B*	*11C*	*11D*	*11E*
Actual responses	164,00	7,00	28,00	61,00	62,00
FV	0,52	0,02	0,09	0,19	0,20

ITEM #14

This item was easier than #11, it tested the same skill, and although its FV score was a little above our target (0.74), it was quite successful. This was mainly because DI value it generated was 0.51, which is exceptional good.

Table 15—12 Sub-test in02 Items #14 Question stem and options (a ✓ indicates the key response)

14. Before deciding to marry Crown Prince Naruhito, Miss Owada was ...

A an anti-feminist. ____

B a traditional Confucian. ____

C a model of selfishness. ____

D a "yellow cab". ____

E a career woman. ✓

The spread of incorrect response was not evenly balanced across the four distractors, but three of the options were chosen by a number of candidates. It would seem that the success of these incorrect options lies in the need for candidates to have a good grasp on information appearing at different points in the text.

Table 15—13 Sub-test in02 Item #14 Facility values for all options (shading indicates the key response)

Item	*14A*	*14B*	*14C*	*14D*	*14E*
Actual responses	1.00	12.00	42.00	28.00	233.00
FV	0.00	0.04	0.13	0.09	0.74

ITEM #13

By comparison with item #11, this is more difficult, and does not discriminate as well. The task invites candidates to interpret the text, and to use their reading skills in order to extract implicit content (Table 15—12).

Table 15—14 Sub-test in02 Item #13 Question stem and options (a ✓ indicates the key response)

13. Once she becomes a member of the Japanese imperial court Miss Owada will probably...

A be in a position to bring about change. ____

B Represent female independence in modern Japan. ____

C be referred to as being a "yellow cab". ____

D find her life comfortable, but limited. ✓

E follow the example set by the Princess of Wales. ____

Once again, we find a good spread of choice among the options, although one is chosen far less often than the others (Table 15—13). This option, #13C, is the most clearly incorrect, and candidates would be able to deduce this because it is neither implicit nor explicit in the text. The first reference to *"yellow cabs"* appears in the middle of paragraph E (Table 15—1), and it is clearly not referring to Miss Owada. The second reference is in the first sentence of paragraph F, and here the text explicitly denies that that she was ever "such a woman".

The other four options all contain information or ideas which are clearly contained in the text, and about which nothing explicit is said. Candidates would therefore have to read the text very carefully in order to discern which is correct.

Table 15—15 Sub-test in02 Item #13 Facility values for all options (shading indicates the key response)

Item	*13A*	*13B*	*13C*	*13D*	*13E*
Actual responses	109,00	40,00	9,00	130,00	26,00
FV	0,34	0,13	0,03	0,41	0,08

ITEM #19

The ability to candidates to deduce word meaning from morphology and context was the target of this item, which was easier than we would have liked, but which achieved the highest DI score of the set of items.

Table 15—16 Sub-test in02 Item #13 Question stem and options (a ✓ indicates the key response)

19. In paragraph C, the phrase "to play the game" is synonymous with ...

A victory. ____
B conformity. ✓
C defeat. ____
D stalemate. ____
E rebellion. ____

The stem refers to a sentence in paragraph C, which reads:

> *By refusing to play the game, by insisting on her freedom, she [the Princess of Wales] has jeopardised an ancient institution.*

In order to deduce the correct answer candidates would need to understand the relationship between the first two phrases. Many of the weaker candidates chose option #19E, which is in line with the meaning of the second phrase, but not the target of the item. In this instance, the spread of responses across the option was less evenly balanced than in the earlier items.

ITEM #20

The grammatical cohesion of the text was the objective of this item, in which candidates were asked to identify the referent of *it* from the sentence quoted (Table 15—17). The item was relatively easy, with an FV of 0.75, but it too discriminated well: DI 0.52.

Another positive attribute of the performance of this item is the fact that the options all attracted a number of responses (Table 15—18).

Table 15—17 Sub-test in02 Items #20 Question stem and options (a ✓ indicates the key response)

20. . In the phrase "That it was an act of self-sacrifice ..." in paragraph C, the word "it" refers to ...

A being selfish. ____

B playing the game. ____

C getting married. ✓

D being useful. ____

E pursuing a career. ____

Table 15—18 Sub-test in02 Item #20 Facility values for all options (shading indicates the key response)

Item	*20A*	*20B*	*20C*	*20D*	*20E*
Actual responses	16.00	37.00	238.00	5.00	21.00
FV	0.05	0.12	0.75	0.02	0.07

ITEM #17

The item that performed worst on this paper was #17. This item was an attempt to test candidates' ability to interpret the text by reading between the lines. The sentences referred to in the stem appear in paragraph E:

> *One could call this selfish and sad. One could also say that society has not kept pace with its most intelligent and independent women.*

As the reader can see, only two of the options have naything to do with the text, and both of these actually quote phrases from the original. The other three options are asking candidates to perform some mental juggling around the first two options. Judging by the poor responses (FV 0.12; DI 0.13) it seems likely that the item was poorly conceived and that format of the distractors "tricked" candidates as much as the item itself contained any difficult (Table 15—15).

Table 15—19 Item #17 Question stem and options (a ✓ indicates key response)

17 In paragraph E, from the sentences "One could call this ... independent women." we can assume that the author of the article does himself believe ...

A. that this is "selfish and sad". ____

B. that "society has not kept pace". ✓

C. neither 'A' nor 'B' is true. ____

D. both 'A' and 'B' are true. ____

E. something completely different. ____

Table 15—20 Sub-test in02 Item #17 Facility values for all options (shading indicates the key response)

Item	*17A*	*17B*	*17C*	*17D*	*17E*
Actual responses	16,00	37,00	238,00	5,00	21,00
FV	0,05	0,12	0,75	0,02	0,07

15.3 Descriptive statist ics

This test produced one of the more normal frequency distribution curves (Figure 15—5). The most notable characteristics of the curve are the fact that the range from lowest to highest score is relatively small, from 5 to 16, and that the mode is 10.5, because there are two equally high peaks, at 10 and 11. This "bunching" of all of the candidates suggests a high degree of homogeneity in the group, and backs up the evidence that relatively few items discriminate well between stronger and weaker candidates. The level of skewness is small (Table 15—21) and suggests that the test was a little easy. The fact that no candidate was able to obtain maximum marks is probably due to defects in the item writing, rather than in candidate ability.

15.4 Reliability

As we have discussed earlier, the question of statistical reliability is one we would have hoped to resolve in the process of trialling and using earlier tests. Our target is to achieve high levels of reliability on the major sub-tests used for full scale testing (≥0.9). In order to calculate the coefficient of reliability we

Figure 15—5 Sub-test in02 Frequency distribution of raw scores out of 20

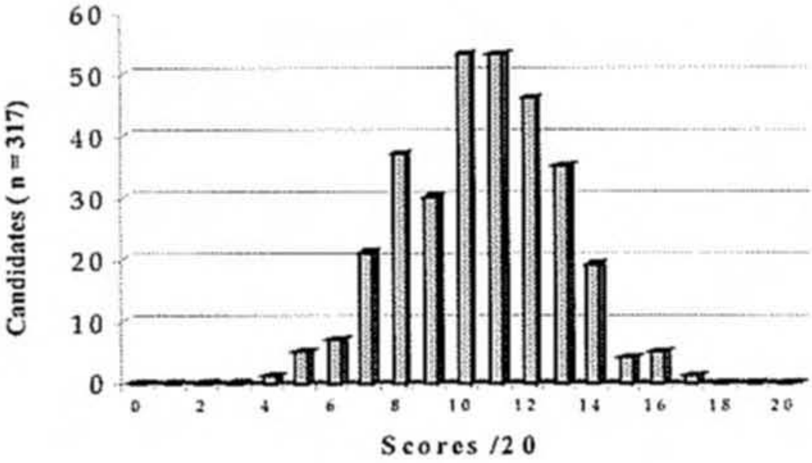


Table 15—21 Descriptive statistics 20 items

Mean	10.50
Median	11.00
Mode	10.50
Standard error of measurement	1.55
Standard deviation	2.35
Skewness	(0.08)
Level of confidence: 95%	0.26

have adopted Bachman's suggestions of using two different split-half formulae. We have calculated the coefficient with each of these twice, on each occasion using a different method of selecting the rank order of the items on which to divide the test into two supposedly equal halves. The resulting overall average for Sub-test in02 was extremely disappointing. A coefficient of 0.395 effectively means little reliability.

Table 15—22 Calculations of average split-half reliability coefficients

	FV rank-order	DI rank-order	Average
Spearman-Brown			
Rxx1	0.315	0.319	0.317
2rhh1	0.374	0.380	
Rhh1	0.187	0.190	
1+rhh1	1.187	1.190	
Guttman			
Rxx1	0.418	0.527	0.473
S2h1	2.041	2.034	
S2h2	2.331	2.038	
S2x	5.529	5.529	
Average	0.367	0.423	
Overall average	0.395		

From a technical point of view however, by applying the SPEARMAN-BROWN PROPHECY FORMULA (Table 15—23) we have calculated that were we to extend the text by increasing the number of similar items from 20 to 34, then we would achieve a coefficient of 0.948, which would be perfectly acceptable. However, as we have said in our analysis of other sub-tests, the reality of producing another 14 similar items, whether based on the same text or not, is not necessarily so straightforward.

Table 15—23 Spearman-Brown Prophecy Formula

N	1.70
r_{tt}	0.395
r_{ttn}	0.948
Items	34

The level of the Standard error of measurement is, unfortunately, high. Candidates' true scores may have differed by more than ±1.95 from their observed scores.

Table 15—24 Standard error of measurement

Standard deviation	2.51
Reliability	0.39
1-r	0.61
Square root of (1-r)	0.78
Standard error of measurement	1.95

15.5 Conclusions

One of the basic principles on which our sub-tests have been designed is that of the discrete analysis of reading skills, hence the use of Munby's taxonomy (1978). What we have learned from the process of analysis of the candidates' performance is that the choice of an individual skill as the target of our test item does not mean that all candidates will use the target skill, and only the target skill, in order to respond. The reading process is clearly more complex, and our discussion of the responses to individual items informs our understanding of this complexity, without being able to reveal the empirical base. However, this does not invalidate the decision to focus items on individual skills. It merely means that we cannot expect human beings to "jump through hoops" in the way in which we have introspectively decided that they should.

Despite this apparent complexity, we can say that the items which have focused on skills to do with understanding information explicitly or implicitly stated in the text, do appear to be sufficiently difficult, and to discriminate sufficiently well between candidates, for this combination to be re-used in further tests. In addition, we have seen two other skills produce very high discrimination indices, in spite of being relatively easy. These items tested skills 19 and 32, respectively, both of which are skills that we have identified in other sub-tests as being particularly useful. These items have combined with the MCQ format to provide good discriminators at the lower level of the ability range.

One clear conclusion from this sub-test is that the time factor can interfere with candidate performance both when they have too little time, and when they have too much. Our sub-tests have been designed to ensure that all candidates were able to complete the paper comfortably within the time limit. This has meant an allowance of two hours, in which to respond to twenty items: six minutes per item. This was a conscious decision taken because of the experimental nature of the test, and in the face of the advice found in the literature (Madsen 1983) that no more than two minutes per item is needed.

Table 15 — 25 Sub-test in02 Item by item summary of performance (shading indicates those items that achieved both target performance levels)

Item	Task type	Skill (Munby 1978:126-131)	FV	DI
#1	Matching titles to paragraphs	Scanning to locate specifically required information on one point, involving a complex search (Skill 46.2)	0.91	0.09
#2		Idem	0.74	0.09
#3		Idem	0.77	(0.13)
#4		Idem	0.67	0.18
#5		Idem	0.28	0.31
#6	True/false/ no evidence	Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)	0.71	0.31
#7		Understanding relations between parts of text through the grammatical cohesion device of reference (Skill 32.1)	0.63	0.06
#8		Understanding explicitly stated information (Skill 20)	0.63	0.22
#9		Idem	0.56	0.21
#10		Idem	0.64	0.05
#11	5-foil MCQ	Idem	0.52	0.49
#12		Understanding information in the text, not explicitly stated, through making inferences; figurative language (Skill 22.1 & 22.2)	0.36	(0.06)
#13		Idem	0.41	0.33
#14		Understanding explicitly stated information (Skill 20)	0.74	0.51
#15		Understanding information in the text, not explicitly stated, through making inferences; figurative language Skill 22.1 & 22.2)	0.31	(0.03)
#16		Idem	0.46	0.17
#17		Interpreting text by going outside it, using exophoric reference "reading between the lines" (Skill 34.1 & 34.2)	0.12	(0.13)

Item	Task type	Skill (Munby 1978:126-131)	FV	DI
#18		Understanding information in the text, not explicitly stated, through making inferences; figurative language (Skill 22.1 & 22.2)	0.48	(0.17)
#19		Deducing the meaning and use of unfamiliar lexical items, through understanding word formation (Skill 19.1) and context (Skill 19.2)	0.87	0.59
#20		Understanding relations between parts of text through the grammatical cohesion devices (Skill 32)	0.75	0.52

16 SUB-TEST ES04

16 SUB-TEST ES04

16.1 TEXT
16.2 TASK TYPES
16.2.1 *Four option Multiple choice items*
16.2.1.1 Facility values
16.2.1.2 Discrimination indices
16.2.2 *Discussion*
Items #6 and 10
Item #9
Items #3 and 4
Item #1
Item #2
Items #7 and #8
Item #5
16.2.3 *Error identification and correction*
16.2.3.1 Facility values
16.2.3.2 Discrimination indices
16.2.4 *Discussion*
16.2.4.1 Error identification items
16.2.4.2 Error correction items
16.3 DESCRIPTIVE STATISTICS
16.4 RELIABILITY

THIS SUB-TEST was administered to a sample of 55 candidates for entry to the first degree program in Translation and Interpreting. They formed part of the potential first intake to the private University Alfonso X AEl Sabio@, located near Madrid, in July 1994. The test, the same model as that being used in Granada, was designed, written and administered as part of a research contract between the two universities (University of Granada, *Agencia de Transferencia de Investigación* contract Nº: 440a. Principal researcher: Dorothy Kelly. Associate Researcher: Bryan Robinson).

Due to the nature of the contract, and the potential difficulties involved in a long-distance follow-up study of results, it was decided not to proceed beyond the entry test stage. As the number of candidates involved in the sample was above our statistical minimum, although it was low by comparison with Granada, the results are included here for two reasons. The first of these concerns the nature of the test itself and the changes made in the test specification with respect to 1993. The second is to do with the political decisions and subsequent changes that had taken place in Granada, and which affected the, writing, registration for, and administration of the tests. The fact is that entry tests are instruments of socio-political control — society's gatekeepers — and their use can never be considered in abstract.

In July 1994, the test specification was altered to introduce the use of the error identification and error correction items, which we study in more detail below. These items constitute an attempt to include objective activities that resemble one part of the translation process tested through the traditional synthetic summary writing exercise, which we discussed earlier. Essentially, they represent a step towards the use of a written production exercise, but they maintain a high degree of objectivity of marking. In this sub-test and in in03 used in the corresponding session in Granada, this task type was used — without trialling — for the first time.

The second change, to which we referred earlier, involved the registration of candidates for the entry test, the makeup of the Examination board, and the preparation of the test papers.

For the first time, prospective candidates for the entry test were to be charged a registration fee. This decision was taken to bring the entry test in line with other tests, and the immediate beneficiary of this fee was to be the newly established Faculty. Members of the Examination board received no payment for the 1994 entry test. As well as this, the Chair of the Examination board passed from the Dean of Faculty to a Vice-Dean who was not a staff member of the Department of Translation and Interpreting. Furthermore, the Department of Spanish Philology took on the role of preparing the LA sub-test. Previously, members of this department had declared themselves against the administration of an entry test in Language A on the grounds that candidates who had just completed their University entrance examinations had been fully examined in the pertinent skills. Responsibilities for the other sub-tests were divided up between the Department of Translation and Interpreting and the corresponding Philology departments. The former took responsibility for the reading skills tests, and the latter for the listening skills. Membership of the Board was based on equal representation by department, and included representatives of the Administration and Services staff.

The sub-test was administered in Madrid by the two researchers involved in the research contract with the support of staff of the Alfonso X University. It was marked immediately after the session, and scores were later published. Details of candidate performance were reported directly to the University (See Appendices).

16.1 Text

The text (Pérez Reverte, Arturo; *El Ideal Revista Dominical*, 1994) was presented in word-processed format with the paragraphs identified by numbers (Table 16—1). For the multiple-choice items, only the first three paragraphs were required. These added up to some 440 words. The fourth paragraph was used for the error identification and error correction items. This format was preferred to the photocopied texts used in earlier sessions because it enabled us to present the error identification and correction items in two columns.

The topic was a personal commentary on present-day reading habits and it was thought to require no specialist knowledge, and that it was unlikely to favor any particular group of candidates. The text was chosen and items written by Dr Francisco Salvador Salvador, a member of the Department of Spanish Philology and the items were written with a clear profile of the potential candidate in mind. The author has extensive experience of testing through the University entrance examination, and has regularly participated in examination boards for the University entrance examination — *selectividad*. The sub-test was neither trialled nor revised by other specialists prior to its administration, and non-specialists who adhered strictly to the markscheme provided by the author carried out marking.

Table 16—1 Sub-test es04 Exam text *Miedo a los libros* (Pérez Reverte, Arturo; *El Ideal Revista Dominical*, 1994)

(1) El prestigio de la lectura como placer ha ido aumentando a medida que crecía el prestigio de la imagen como fuente principal de diversión. Se ha fomentado la idea de que la letra es causa de aburrimiento, mientras que cuanto llega a través de la imagen es necesariamente ameno. Esta teoría ha creado, como es sabido, un hábito de pereza mental, pero también una aceptación de lo pesado como si fuese ligero. Según esta actitud, cualquier espacio de la televisión entretiene más y mejor que la más entretenida novela de aventuras. Se equivoca la masa a tal respecto. O, peor aún, no se concede a sí misma la posibilidad de rectificar abriendo un libro de vez en cuando. A menudo he citado el caso del adolescente que descubrió cuán amena puede resultar la lectura de La Odisea si se aprende por voluntad propia y sin la obligación de aprobar un examen. Al igual que él, muchas personas que retroceden ante la lectura se asombrarian al descubrir que muchos títulos en apariencia rimbombantes de nuestro pasado cultural son mucho familiar implica necesariamente torpeza, cretinez y vulgaridad.

(2) Paralelamente a estos fenómenos típicos de las normas de coacción que rigen la segunda mitad del siglo, asistimos en las librerías a un renacer de la literatura de aventuras. Con el propósito de servir a los intereses y gustos de los más jóvenes, se reeditan títulos míticos del género, combinando el colonialismo poético de Kipling con las profecías, a la larga sensatas, de Julio Verne. Claro que no todo el monte es orégano, aún pareciéndolo. Algunos autores favoritos de la literatura aventurera han desaparecido de los catálogos, cuando años atrás tuvieron el poder de cautivar a toda una generación de jovencitos. ¿Quién lee hoy las aventuras de Tarzán, héroe que parecía imbatible? ¿Quién recuerda a los centauros de Karl May o Zane Grey?

(3) No pertenezco al grupo de ilusos inclinados a exigir que todos sus vecinos de escalera hayan leído el Ulises o La Montaña mágica. Para que muchos vecinos perdiesen el miedo al libro y se apartasen de la tiranía de

más amenos que cualquiera de estos telefilmes americanos que transcurren invariablemente en una comisaría y abundan en personajes y situaciones repetidas hasta la exasperación. Y no hablemos ya de las innúmeras comedietas de corte casero empeñadas en demostrar que la vida la imagen no harían falta metas tan elevadas: bastaría con ayudarles a recobrar la pasión que en otro tiempo le inspiraban otros títulos aparentemente simplones, de los que solíamos decir que "se leen de un tirón". Cuenta mucho, de todos modos el cambio de época, con la consiguiente transformación de la sensibilidad.

16.2 Task types

16.2.1 *FOUR OPTION MULTIPLE CHOICE ITEMS*

This test saw the use of four-option multiple choice items, instead of the five-option items that had been written for earlier sub-tests. The change was introduced for two reasons. Firstly, because of the poor results produced by the five-option items and, secondly in an attempt to reduce the level of difficulty of the MCQ items which were now the first set on the paper.

The skill selection was based on the same specification as that used in earlier tests, and this had been published by the University along with registration details.

Table 16—2 Sub-test es04 Items #1-10 Target skills (Munby 1978:126-131)

Item	*Target skills*
#1, #4, #6, #10	Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)
#2	Recognizing specific literary figures
#3, #7, #8	Deducing the meaning and use of unfamiliar lexical items, through contextual clues (Skill 19.2)
#5	Understanding relations between parts of text through lexical cohesion devices of pro-forms/general words (Skill 30.7)

Item	*Target skills*
#7	Understanding relations between parts of text through lexical cohesion devices of antithesis (Skill 30.4)
#9	Interpreting text by going outside it, integrating data in the text with own experience or knowledge of the world (Skill 34.3)

Item #2 was not based on a skill given in the specification, nor does this appear as such in Munby (1978:126-131). It was included in the test paper by the author in the belief that this was a skill candidates who had recently completed Spanish secondary schooling would and should be able to demonstrate. The item was included in the final count, and scored as normal.

16.2.1.1 *Facility values*

Overall, this set of items failed to test candidates adequately. Given that this was a test of the candidates' first language, we would have expected a range of Facility values near or above the target maximum of 0.7. In fact, only one item proved easier than required, and half of the set were more difficult than the minimum. Only items #3, #4, #6, and #10 came within our range. Three of these focussed on the same target skill (Figure 16—1).

16.2.1.2 *Discrimination indices*

The performance of this set in terms of discrimination was also disappointing. Only three items achieved the minimum, two of which also had FV scores within our range (Figure 16—2). These items, #6 and #10 will be discussed in more detail later. Item #9 also discriminated to a limited extent, but this item was difficult. All of the other items were below the target minimum, and one, #5, produced a negative figure.

16.2.2 *DISCUSSION*

Some of the causes for the poor performance of this sub-test will be seen to have derived from the preparation of the paper itself. Others are due to the difficulties inherent in the uncoordinated writing of test papers, when neither peer evaluation nor trialling is carried out. Others have to do with aspects of writing MCQ items. In our discussion, we will make use of the FV scores for all of the options in order to contrast performance.

Figure 16—1 Sub-test es04 Items #1-10 Facility values

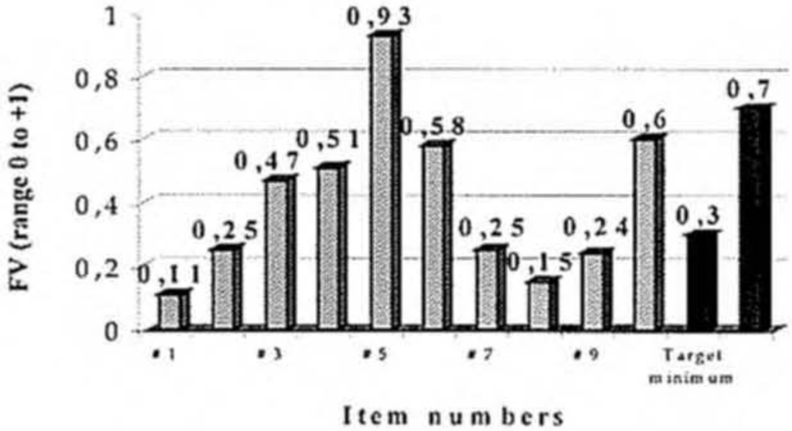
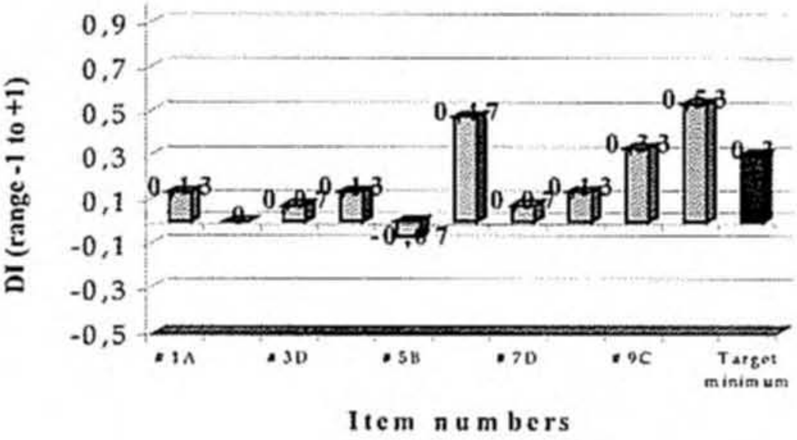


Figure 16—2 Sub-test es04 Items #1-10 Discrimination indices



ITEMS #6 AND #10

These two items focussed on the skill of inferencing, as shown in table 16—3. In the case of #6, we can see that all of the options offered were plausible, but only one was clearly correct given the context of the phrase.

Table 16—3 Sub-test es04 Item #6 Question stem and options (a ✓ indicates the key response)

6 "... las profecías, a la larga sensatas, ..." (párrafo 2). La expresión subrayada se podría sustituir por ...

A. Cuando ha pasado un tiempo. ✓

B. a su debido tiempo. ______

C. en su momento. ______

D. Siempre. ______

Given that the distractors do appear plausible, we would expect an even spread among the alternatives chosen by candidates. In this case, that would mean some eight responses for each of the alternatives, and FV scores of around zero. 15. The spread was uneven, although all options did attract some responses.

Table 16—4 Sub-test es04 Item #6 Facility values for all options (shading indicates key response)

Item	*#6A*	*#6B*	*#6C*	*#6D*
Actual responses	32	14	3	7
FV	0.58	0.25	0.05	0.13

Candidates without actually reading the passage could reject alternatives #6C and #6D, whereas #6B probably needs reference to the text to ensure it is incorrect.

In the case of item #10, which appears below, we find a similar scenario. The skill targeted is also inferencing, and the spread of responses attracted by the distractors could be expected to be slightly lower or the same as for item #6. However, the first two are easily ruled out as each is based around a word that carries elements of negation in its meaning: *carencia* and *ausencia*. The decision then is limited to a choice between #10C and #10D, and the only real difference is in the strength of the reference. Candidates would, again, probably be able to make a decision about this item without referring to the text, and choose #10C, because it represents a more forceful alternative, more in line with the use of *tirania* in the stem.

Table 16—5 Sub-test es04 Item #10 Question stem and options (✓ indicates key response)

1 " ... la tiranía de la imagen ..." (párrafo 3) se refiere a la
0 ...

A. Carencia de otros intereses visuales. ______
B. Ausencia de otros medios. ______
C. Atracción absoluta hacia el cine y la TV. ✓ ______
D. Influencia de los medios audiovisuales. ______

ITEM #9

This item proved difficult with an FV of 0.24, but it was a good discriminator, with a DI of 0.33. The inclusion of the item, based on the candidates' knowledge of the texts quoted and their ability to identify them as contemporary literature, or choose the correct alternative by elimination, corresponded to the test author's knowledge of the potential candidature. Superficially, all of the alternatives could have been correct, and the spread of responses was good.

Table 16—6 Sub-test es04 Item #9 Facility values of all options (shading indicates key response)

Item	*#9A*	*#9B*	*#9C*	*#9D*
Actual responses	9	15	13	18
FV	0.16	0.27	0.24	0.33

Ideally, each alternative would have attracted 14 responses, and thus achieved a FV of 0.25. It is likely that candidates who opted for #9D did so because they confused *Ulysses*, by James Joyce, with the Greek classic, and assumed that *La Montaña Mágica* was neither Greek nor a classic. Again, in the answering of this item we see that there was much room for candidates to score the points by drawing on their general cultural background knowledge and without having recourse to their reading skills.

ITEMS #3 AND 4

The other two items that achieved average FV scores, were #3 and 4, with 0.47 and 0.51, respectively. Each of these focussed on a different skill. Item #3 was based on

> *Deducing the meaning and use of unfamiliar lexical items, through contextual clues (Munby 1978:126-131 Skill 19.2).*

Effectively the stem asked candidates to choose a synonym for the word underlined from those given in the options.

Table 16—7 Sub-test es04 Item #3 Question stem and options (✓ indicates key response)

3 "... muchos títulos en apariencia rimbombantes ..." (párrafo 1) ¿qué adjetivo estaría más próximo a la palabra subrayada?

A. Sonoro. ____

B. Rumboso. ____

C. Glorioso. ____

D. Llamativo. ✓

One obvious distraction in the wording of this item lies in the fact that the adjectives used in the options are all presented as masculine singular forms, when the focus of the stem is on a masculine plural form. This type of incoherence is not necessarily going to cause difficulties to candidates, but it might, and given the nature of the item it constitutes an unnecessary distraction.

Table 16—8 Sub-test es04 Item #3 Facility values of all options

Item	#3A	#3B	#3C	#3D
Actual responses	15	9	5	26
FV	0.27	0.16	0.09	0.47

As can be seen, all of the alternatives attracted some responses, although the spread was not balanced. As 26 candidates chose the key response, giving this an FV of 0.47, the ideal FV score for each of the others would have been around 0.17. In this instance, alternative #3A, perhaps the least unusual of the distractors, has attracted rather more than its fair share.

Item #4 focussed on

> *Understanding information in the text, not explicitly stated, through making inferences (Munby 1978:126-131 Skill 22.1).*

However, in fact it presented three distractors that left little choice to candidates. What is surprising is that it should have achieved such a low FV score, given the manner in which the options are presented. In an MCQ item, any clear indication that one of the options is in some way different to the others marks the option in question as either clearly the key response, or completely wrong. In this case, without even reading the text, candidates would have been able to judge that option #4A is correct. They could do this, firstly because it is clearly different to the other three, and secondly from their knowledge of morphology, which would lead them to associate the form *–ietas*, in the question stem, with the adjective *despectivo* in the option.

Table 16—9 Sub-test es04 Item #4 Question stem and options (✓ indicates key response)

4 "Comedietas de corte casero ..." (párrafo 1). Con esta expresión el autor se refiere ...

A. al uso despectivo del término comedia. ✓

B. a la Comedia de costumbres. ____

C. a la Comedia de carácter. ____

D. a la Comedia familiar. ____

Having said that, we need now look at actual performance. In fact, the item was more difficult than we might have predicted, and the range of FV scores achieved by the options was quite high. Why did option #4D attract 22 responses? Quite possibly because candidates did actually read the item and try to respond with reference to the text.

Table 16—10 Sub-test es04 Item #4 Facility values of all options

Item	*#4A*	*#4B*	*#4C*	*#4D*
Actual responses	28	5	0	22
FV	0.51	0.09	0.00	0.40

If we now look briefly at items #1, 2, 7 and 8 we may find some sort of explanation for their difficulty other than any connection with the target skills on which they focussed. Each of these items was based on a different skill, and each presented different difficulties of a technical nature, which might have been eradicated had the sub-test been given to one or more colleagues for revision.

ITEM #1

This item proved very difficult, and was deficient in terms of discrimination. Why was this the case? As the first item on the paper, and in line with the generally established policy, it would have been best to offer candidates an item which was at least of average difficulty.

The difficulty of the item lies in the fact that the affirmation of the first sentence of the text does not appear at all logical unless and until it is seen in the context of the full passage.

Table 16—11 Sub-test es04 Item #1 Question stem and options (✓ indicates key response)

1 Según el texto, la idea de que la lectura provoca aburrimiento, (párrafo 1) ¿qué hábito ha creado entre la gente?

A. Potenciar la cultura de la imagen. ____

B. Dificultad de concentrarse. ✓

C. Acostarse más temprano. ____

D. Ver más la TV. ____

Again, options #1C and #1D effectively cancel each other out, as there is a degree of opposition in the content of each. In addition, option #1B seems illogical if we carry out a first superficial reading of the text. Moreover, it is the only one of the four options that uses a noun as the headword rather than a verb. Option #1A seems both the most logical, if we relate the text to our knowledge of the world, and the most accurate unless we spend a lot of time on the text. However, as we have shown in, the key response was #1B. The justification for this, in the view of the author, is to be found in Paragraph 1, in the phrase

...un hábito de pereza...

considered synonymous with option #1B.

Candidates evidently did not see this, and the choice of the majority was for option #1A.

Table 16—12 Sub-test es04 Item #1 Facility values of all options (shading indicates key response)

Item	*#1A*	*#1B*	*#1C*	*#1D*
Actual responses	29	6	0	20
FV	0.53	0.11	0.00	0.36

Furthermore, option #1D gained a significant number of responses, with the key response attracting only six candidates.

In our analysis of the paper carried out after the marking session, we gave some thought to the question of whether or not this item was valid. It could not be said that the key response was incorrect, but it was clearly a very difficult item. In a revised version of the test this item would either have been replaced or made easier. As things stood, the item was counted as valid for the sample taking the test, and the markscheme was applied as it stood.

ITEM #2

A superficial analysis of this item from the point of view of a test task shows that it should function adequately. It is only when the item is compared with the test specification that we see it is based on the knowledge of a series of concepts that are external to the test. Whether or not we can assume that candidates know these items is not relevant here. The importance of this lies in the fact that the item contravenes the test specification as it depends not on the ability of candidates to demonstrate their use of reading skills, but on their content knowledge of rhetorical figures. It is a clear example of an item that fails to meet criteria of construct validity.

Table 16—13 Sub-test es04 Item #2 Question stem and options (✓ indicates key response)

2 "Cualquier espacio de la TV entretiene más y mejor que la más entretenida novela de aventuras ..." (párrafo 1) ¿qué figura retórica parece que emplea el autor en esta frase?

A. Anáfora. ____

B. Paradoja. ✓

C. Anacoluto. ____

D. Redundancia. ____

Technically, the item is presented well. Each of the alternatives is a single noun, and each belongs to the same semantic category. The flaw lies in the underlying assumption about the content of the item and the test specification. However, the item did not serve any significant purpose in the test, as the FV score was low, 0.25, and discrimination was zero. Effectively it was an item wasted.

ITEMS #7 AND #8

These items have one target skill in common, namely that of

> *Deducing the meaning and use of unfamiliar lexical items, through contextual clues (Munby 1978:126-131 Skill 19.2).*

The difference between them lies in the more complex mechanism of item #7, which requires the use of a second skill in order to handle the stem and options. This second skill —

> *Understanding relations between parts of text through lexical cohesion devices of antithesis (Munby 1978:126-131 Skill 30.4)*

— which is closely related in the first in that it also deals with an aspect of lexis makes the item seem more difficult. The reality of the results, however, is that both items are difficult, with FV scores of 0.25 and 0.15 respectively. Neither item discriminates to a significant degree (#7 DI 0.07; #8 DI 0.13).

Table 16—14 Sub-test es04 Items #7-8 Question stem and options (✓ indicates key response)

7 "... tuvieron el poder de cautivar a toda una generación ..." (párrafo 2). ¿Cuál sería el verbo más exactamente contrario de cautivar?

A. Desencantar. ____

B. Aborrecer. ____

C. Rechazar. ____

D. Repeler. ✓

8 "Quién recuerda a los centauros de Karl May ..." (párrafo 2). Con qué palabra o palabras podríamos sustituir el término "centauros" según el sentido del texto.

A. Vaqueros (cow-boys). ✓

B. Jinetes audaces. ____

C. Policía montada. ____

D. Caballeros. ____

Why have these items failed to discriminate between candidates? In the case of #7, the answer probably lies in the extreme difficulty of the question stem, and in the fact that the four alternatives are very close in meaning. That option #7D should be correct is difficult to justify.

Table 16—15 Sub-test es04 Item #7 Facility values of all options

Item	*#7A*	*#7B*	*#7C*	*#7D*
Actual responses	28	9	3	14
FV	0.51	0.16	0.05	0.25

The FVs scored by all of the options show that #7A was the most popular by far. The very nature of the semantic proximity between #7A and #7D makes it likely that, had a prior analysis of results been possible, the key response for this item would have been changed.

Similar problems emerge with item #8. The key response is the third most popular, and it is really only #8C which can be easily discounted. Of the apparently correct options #8A is either the most likely, or the least likely, because it includes the loan word *cow-boys*, which candidates may have rejected precisely because it contains an English word. Again, the item seems to have a number of intrinsic, technical flaws, which would almost certainly have led to its revision, had the sub-test be trialled.

Table 16—16 Sub-test es04 Item #8 Facility values of all options

Item	*#8A*	*#8B*	*#8C*	*#8D*
Actual responses	8	29	2	16
FV	0.15	0.53	0.04	0.29

ITEM #5

The only item we have yet to discuss in this set is #5. This is designed to focus on the skill of

> *Understanding relations between parts of text through lexical cohesion devices of pro-forms/general words (Munby 1978:126-131 Skill 30.7)*

In fact, it seems to have a double focus, as did #7, because a further skill is required in order to manage to relate the stem and options to the item. To a certain extent, candidates must be able to interpret the text

> *By going outside it, "reading between the lines" (Munby 1978:126-131 Skill 34.2)*

Without this ability, they are unable to discern which of the options is correct: the more obvious superficial reference based on the association of *titulos míticos* with *dioses griegos* in #8C, or the key response of #8B.

Table 16—17 Sub-test es04 Item #5 Question stem and options (✓ indicates key response)

5 "... se reeditan titulos míticos del género ..." (párrafo 2). Con esta frase el autor se refiere a los libros ...

A. de profundo sentido religioso. ____
B. más conocidos. ✓
C. de dioses griegos. ____
D. más consultados por los especialistas. ____

If we refer to the FV scores for all of the options, we find that the item produced no difficulties at all, with the highest FV score of the set at 0.93, making it too easy. The item, however, provides the only negative DI score of the section with DI -0.07, which shows that the stronger candidates made mistakes by choosing #5C. Options #5A and #5D attracted no responses at all, demonstrating their inefficiency as distractors.

Table 16—18 Sub-test es04 Item #5 Facility values of all options (shading indicates the key response)

Item	#5A	#5B	#5C	#5D
Actual responses	0	51	4	0
FV	0.00	0.93	0.07	0.00

16.2.3 *ERROR IDENTIFICATION AND CORRECTION*

These items were seen as a significant improvement in the construct validity of the test.

Table 16—19 Sub-test es04 Error identification and error correction items (Key responses appear in the right-hand column in bold italics. Item numbers have been added to facilitate the discussion.)

A continuación, presentamos el último párrafo del texto; en él se han introducido 10 errores de tipo lingüístico (acentuación, ortografía, morfosintaxis y puntuación). Marcar el punto donde se encuentre una incorrección y proponed la solución correcta en la línea de al lado.

Comprendo (11) ~~que al~~ enfrentar ejemplos de la cultura llamada "superior" contra los de la (13) ~~supcultura,~~ me adentro en una selva tan intricada como la de la Comedia. En cualquier caso, conviene disponer de la suficiente humildad para no despotricar contra las novelas que, (15) ~~segun~~ el tópico, "se leen de un tirón". En realidad, (17) ~~restituyan~~ al lector perezoso sus derechos sobre la imaginación, hoy (19) ~~aletardada~~ bajo los sobornos constantes de la caja tonta. Este esfuerzo es el primero que cabría (21) ~~esigir~~. Sigo recordando el ejemplo de mi amiguito el adolescente (23) ~~homerófilo,~~ salió de La Odisea tan encantado que se puso de inmediato a buscar experiencias (25) ~~parecidas a~~ otros libros. (27) ~~Mientras~~ se perdió quince o veinte concursos de varias cadenas. Y si hizo esto un pobre adolescente despistado, (29) ?han de ~~ser de menos~~ mis vecinos adultos?.

Comprendo (12) ***que, al*** enfrentar ejemplos de la cultura llamada "superior" contra los de la (14) ***subcultura,*** me adentro en una selva tan intricada como la de la Comedia. En cualquier caso, conviene disponer de la suficiente humildad para no despotricar contra las novelas que, (16) ***según*** el tópico, "se leen de un tirón". En realidad, (18) ***restituyen*** al lector perezoso sus derechos sobre la imaginación, hoy (20) ***aletargada*** bajo los sobornos constantes de la caja tonta. Este esfuerzo es el primero que cabría (22) ***exigir.*** Sigo recordando el ejemplo de mi amiguito el adolescente (24) ***homerófilo;*** salió de La Odisea tan encantado que se puso de inmediato a buscar experiencias (26) ***parecidas en*** otros libros. (28) ***Mientras,*** se perdió quince o veinte concursos de varias cadenas. Y si hizo esto un pobre adolescente despistado, (30) ***¿han de ser menos*** mis vecinos adultos?.

Table 16—20 Sub-test es04 Items #11-30 Error identification and error correction Target skills (Munby 1978:126-131 Shading indicates items that are not derived from this source)

Item	*Target focus*
#11-12 #23-24 #27-28	Recognizing the script of a language: understanding punctuation (Skill 17.3) Manipulating the script of a language: using punctuation (Skill 18.3)
#13-14 #15-16 #21-22	Recognizing the script of a language: following grapheme sequences (Spelling system) (Skill 17.2) Manipulating the script of a language: catenating grapheme sequences (Spelling system) (Skill 18.2)
#17-18	Knowledge of irregular verb form
#19-20	Knowledge of vocabulary
#25-26	Understanding relations between parts of a text through lexical cohesion devices of lexical set/collocation (Skill 30.6) Expressing relations between parts of a text through lexical cohesion devices of lexical set/collocation (Skill 31.6)
#29-30	Discrimination between levels of register: identification and substitution of colloquialism

The skills on which items #17-18, #19-20, and #29-30 were focussed do not form part of the specification for the test, nor do they appear in Munby (1978). They are, however, skills that the author of the test assumed candidates should be able to demonstrate.

16.2.3.1 *Facility values*

As was perhaps to be expected with candidates being tested in their first language, only four of the error identification items produced FV scores within our target range, and all of these were in the top half (FV <0.5). Items #11, 19, 23, and 29 were of average difficulty, or easy, while the remainder were easy, or very easy. The corresponding correction items, #12, 20, 24, and 30 also fell within our range, and all of the FV scores were slightly lower. Two other items, #26 and 28, also fell within the target range. The FV scores for the error identification items had both been 0.71, just above our target maximum. All of the other FV scores for error correction were above our range.

16.2.3.2 *Discrimination indices*

The value of these items as discriminators was surprisingly good, given their relative ease. Five error identification and their corresponding five error correction items achieved good DI scores, slightly or considerably above our target. Only one recorded a negative DI score, item #14 (DI –0.33).

16.2.4 *Discussion*

16.2.4.1 *Error identification item s*

First, we will look at the performance of the set of odd numbered items and attempt to draw some conclusions as to the reasons for their performance in terms of both facility and discrimination.

The three items that focussed on spelling, including the use of accents, all performed in a similar fashion. These items were very easy, with FVs over 0.81, and two of them failed to discriminate to any great degree: item #13 had a DI of 0.07, and #15 reached 0.2. However, item # 21 passed our target and could be described as a good measure of discrimination at DI 0.33. The target word in the text, spelt wrongly, was *esigir.*

Another sub-set of items, which we can look at together here, is #11, 23, and 27, all of which deal with questions of punctuation. The first of these was of average difficulty (FV 0.51) but deficient as a discriminator at DI 0.13. The second produced good scores for both facility and discrimination: FV 0.58, also of average difficulty, DI 0.38, a good discriminator. Finally, the third of this set was easy, but it also proved a good discriminator (FV 0.71; DI 0.33).

The remaining four items all focussed on different skills, three of which were outside of the test specification. Item #17 was very easy, but discriminated well. It focussed on the use of an incorrect, irregular verb form: *restituyan*.

Item #19 presented the non-existent *aletardada*, instead of *aletargada*. It could be said that this was also a test of spelling, but essentially candidates who were able to decide on the correction probably needed to be able to do more than just spell. Their capacity to distinguish between the plausible but erroneous was being put to the test.

Item #25 required candidates to identify an incorrect collocation in the context of the passage. This was also easy, with an FV just above our target maximum, but a good discriminator: DI 0.4.

The final item, #29, was based on the analysis of register, and candidates had to identify the colloquialism that appeared in

Figure 16—3 Sub-test es04 Facility values of error identification items

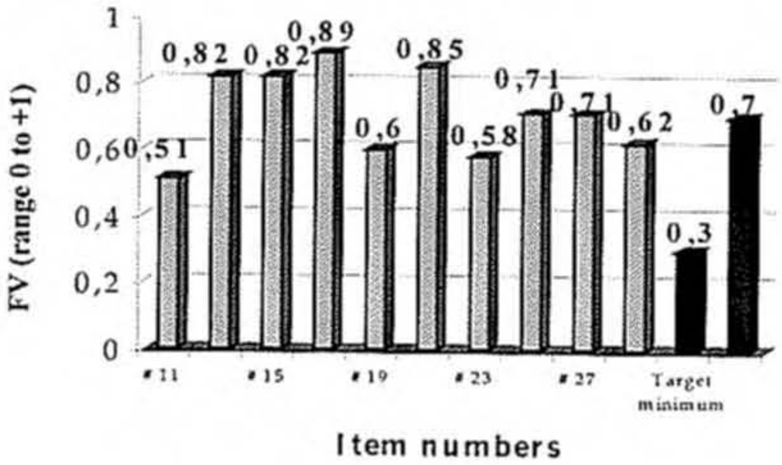


Figure 16—4 Sub-test es04 Facility values of error correction items

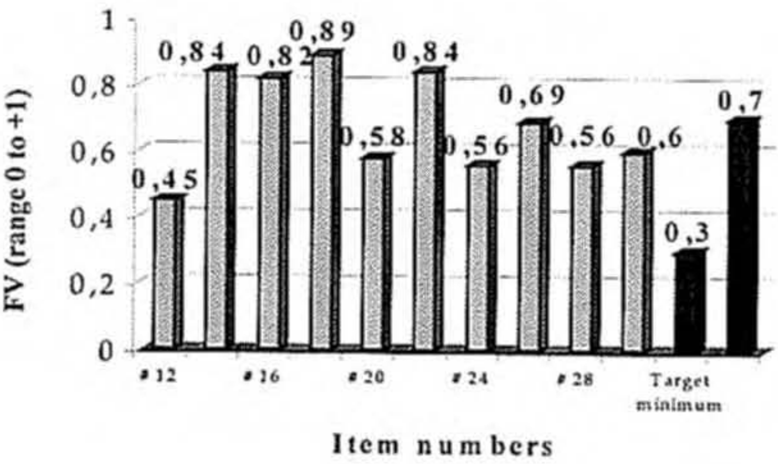


Figure 16—5 Sub-test es04 Discrimintion indices of error identification items

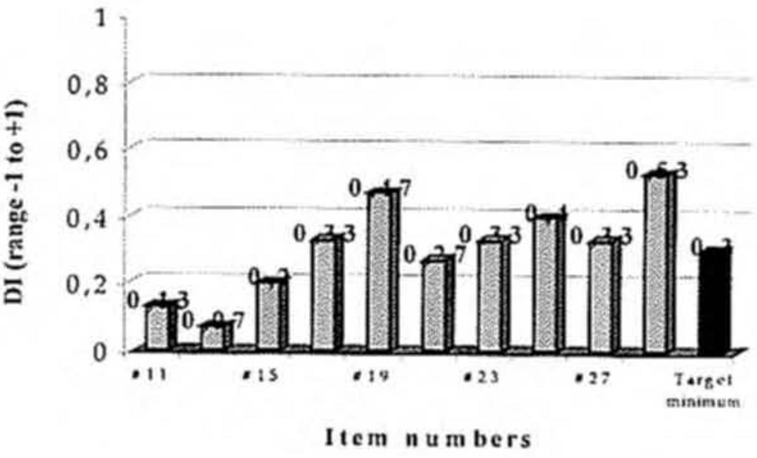
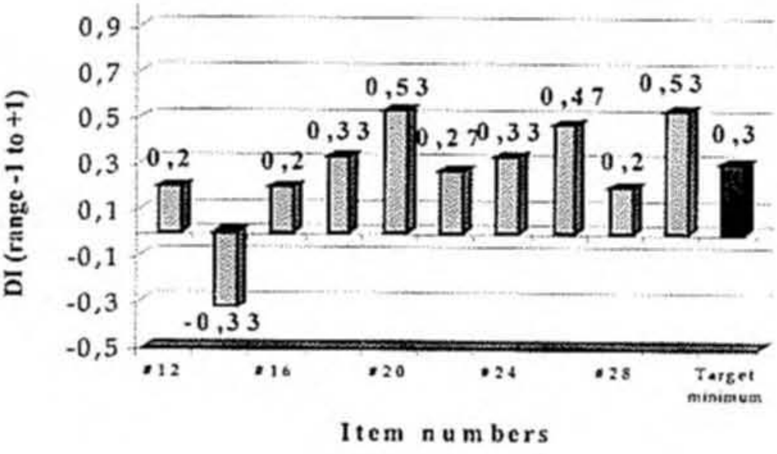


Figure 16—6 Sub-test es04 Discrimintion indices of error correction items



the passage. This item produced a high FV score, 0.62, and the best DI score 0.53, making it an excellent measure of discrimination. The fact that it is also the penultimate item on the sub-test is an added bonus, as it means that candidates have reached the more difficult item when they have been through the paper many times.

16.2.4.2 *Error correction items*

We will now look at the performance of each of the even-numbered items, which involved correcting the errors identified previously. To do this, we will again follow the pattern of looking at the items in sets according to the target skills.

Items #14, #16, and #22 all required candidates to correct errors of spelling that had been inserted into the text. All of them produced high FV scores, showing that they were very easy, and none of them produced a good DI score, although both #16 and #22 were average discriminators at DI 0.2 and 0.27, respectively. The strangest result in this set, and in fact among the 30 items in the sub-test, is that produced by #14. The DI score for this item was –0.33, indicating that it discriminated against the strongest candidates. Candidates chose to change completely the word provided in the text, substituting phrases such as *cultura inferior*. Some of their alternatives were correct, but they were carrying out the wrong task. In this case, as the correction was carried out without MODERATION or INTER-RATER RELIABILITY checks, these candidates were penalized for having provided correct responses.

The set of items in which candidates had to correct errors of punctuation were #12, #24, and #28. The performance of these items was not bad, but it was very unequal in terms of their discrimination. The first item proved of average difficulty and was an average discriminator. The second was of a similar level of difficulty, but a good discriminator with a DI of 0.33. The third was also of average difficulty, although it had a higher FV than the others, but proved an excellent discriminator at DI 0.53.

The final four items, #18, #20, #26, and #30 all produced interesting results, with FV and DI scores that were in line with the corresponding error identification items. In fact, all of them produced good or excellent discrimination indices. Item #18 had a high FV score, making it very easy, but did discriminate, while the others produced FV scores between 0.58 and 0.69, and DI scores of 0.53, 0.47, and 0.53, respectively.

Figure 16—7 Sub-test es04 Frequency distribution of scores for MCQ items

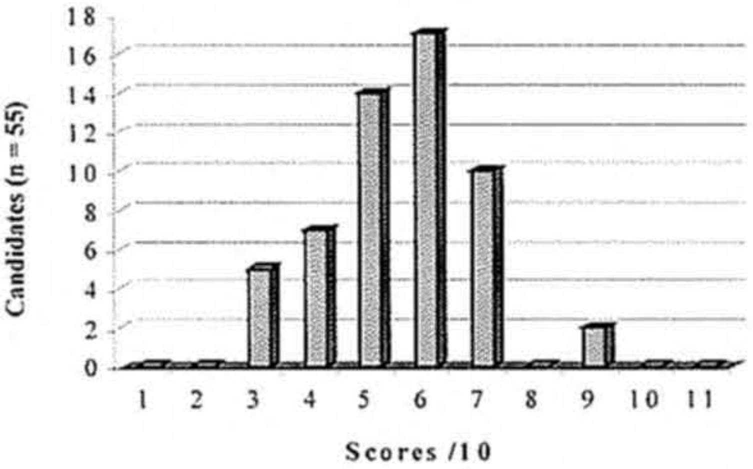


Figure 16—8 Sub-test es04 Frequency distribution of scores for Error identification and error correction items

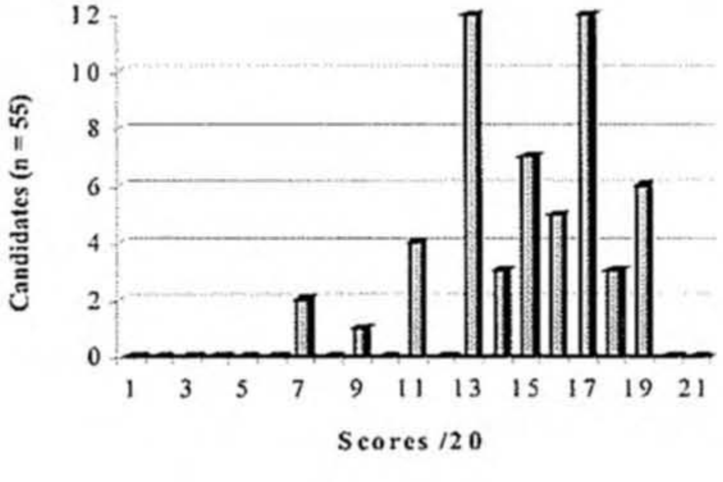


Figure 16—9 Sub-test es04 Frequency distribution of weighted total scores

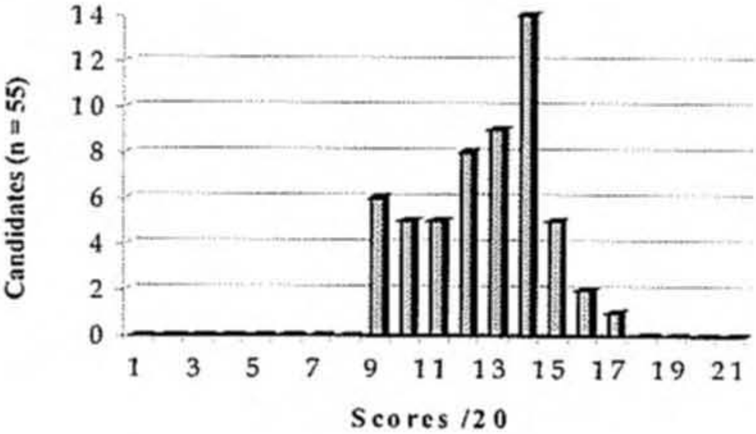
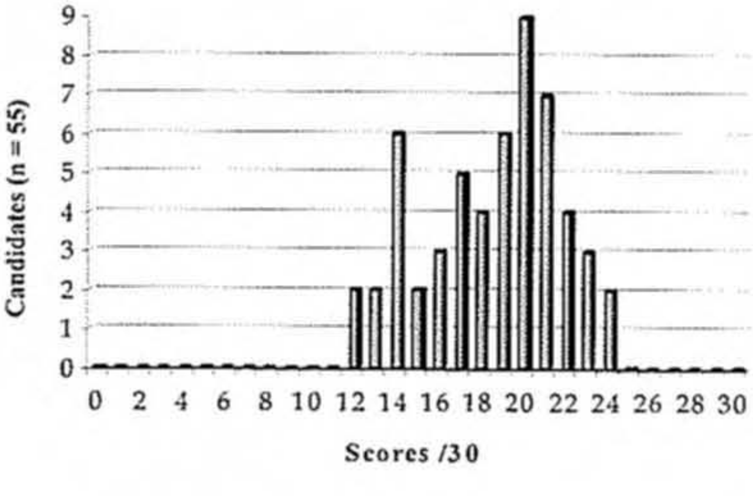


Figure 16—10 Sub-test es04 Frequency distribution of total scores



16.3 Descriptive statis tics

In this section, we are going to look in some detail at the statistics derived from the test and in particular at the comparisons we can make between the different item types. This will involve a discussion of the histograms that represent these visually, and of the tables of descriptive statistics.

The reasons we will carry out this analysis by item type is that the Examination board decided to WEIGHT the three components of the test. In this particular test, weighting the components meant that the raw scores obtained by candidates would not be the same as the final scores.

Table 16—21 Sub-test es04 Breakdown of components to distinguish raw scores from weighted scores

Component	*Maximum raw score*	*Maximum weighted score*
10 MCQ items	10	10
10 Error identification items	10	5
10 Error correction items	10	5

The 10 MCQ items were given their full value with one point each; the 10 error identification and 10 error correction items were given half their value. In this way, the Examination board hoped to compensate for the experimental nature of this component of the test. In fact, when we look at the results, we see that this did influence the final pattern of results. The FREQUENCY DISTRIBUTION of scores for the MCQ items was quite close to being normal (Figure 16—7). The median and mode were the same, and the mean was only slightly lower.

The degree of skewness was minimal (0.11). When we look at the frequency distribution for the error identification and error correction items we see a very different pattern, or to be more precise, a lack of a clear pattern (Figure 16—8). There are two modes, at 14 and 18 out of 20. The median and mean are closer to the first of these, and there is a slight negative skew. When these figures are merged in the weighted scores which were used as a basis for entry decisions, we find that the relative normality of the MCQ items impresses itself on the overall score (Figure 16—9). Essentially, this means that as the aim of weighting the scores was to counteract the more experimental nature of the error identification and error correction items, this was successfully achieved. The frequency distribution histograms display this effect quite clearly.

Table 16—22 Sub-test es04 Descriptive statistics for MCQ items (score out of 10)

Mean	4.51
Median	5.00
Mode	5.00
Standard error	0.19
Standard deviation	1.37
Skewness	0.11
Level of confidence : 95%	0.36

Table 16—23 Sub-test es04 Descriptive statistics for error identification and error correction items (score out of 20)

Mean	13.95
Median	12.00
Mode (bimodal curve)	14.00 & 18.00
Standard error	0.39
Standard deviation	2.90
Skewness	(0.75)
Level of confidence : 95%	0.77

Table 16—24 Sub-test es04 Descriptive statistics for weighted scores (out of 20)

Mean	11.48
Median	12.00
Mode	13.00
Standard error	0.27
Standard deviation	2.00
Skewness	(0.21)
Level of confidence: 95%	0.53

Table 16—25 Sub-test es04 Descriptive statistics for unweighted total scores (out of 30)

Mean	18.45
Median	19.00
Mode	20.00
Standard error	0.43
Standard deviation	3.21
Skewness	(0.34)
Level of confidence : 95%	0.85

To return to the error identification and error correction items for a moment, we will look briefly at some measures that may help us to learn about the comparative performance of these two sets of items. In this way, perhaps we can find further insights into their value, or lack of value. We have calculated the means for each of the sets of scores. The conclusion we reach from these data indicates that for these items, candidates found it as easy to locate the errors, as they did to correct them. However their ability to find errors was marginally more valuable as a means of discriminating between candidates. These figures reveal very little of certainty, and our speculations would need to be corroborated, or not, by a larger.

Table 16—26 Sub-test es04 Comparison of mean Facility values and Discrimination indices for error identification and error correction items

Mean score	*Error identification*	*Error correction*
FV	0.68	0.68
DI	0.31	0.27

16.4 Reliability

One of the advances we hoped to make by introducing the error identification and error correction items was in the area of reliability. The increase in the number of items included in the raw score of the test would, we believed, go some way towards achieving higher coefficients. By using the same procedures as in earlier sub-tests we calculated the average for the split-half coefficients and then went on to apply the Spearman-Brown Prophecy Formula.

Table 16—27 Sub-test es04 Calculations of split-half coefficients of reliability

	FV rank-order	*DI rank-order*	*Average*
Spearman-Brown			
$r_{xx}{}^{1}$	0.501	0.513	0.507
$2r_{hh}{}^{1}$	0.668	0.690	
$r_{hh}{}^{1}$	0.334	0.345	
$1+r_{hh}{}^{1}$	1.334	1.345	
Guttman			
$r_{xx}{}^{1}$	0.488	0.511	0.499
$s^{2}{}_{h}{}^{1}$	4.926	4.274	
$s^{2}{}_{h}{}^{2}$	2.882	3.415	
$s^{2}x$	10.327	10.327	
Average	0.494	0.512	
Overall average	0.503		

The conclusion that we reached is that in this specific sub-test, the first of its type, untrialled, and with a number of errors in the test writing and revision process, we had gained little in terms of reliability. Similarly, the increase in the number of items that would be needed to produce a reliable instrument was +12, which is similar to that we had estimated for earlier tests.

Table 16—28 Sub-test es04 Reliability of a lengthened version

Spearman-Brown Prophecy Formula (42 items)	0.906

The Standard error of measurement for the test, at 30 items, unweighted, was no improvement on that of earlier tests, nor was it significantly worse. In fact, if the final scores had been calculated on the unweighted total it would have been an improvement. We would have been able to assert, with 68% certainty, that any candidate's true score lay within a range of their observed score ±1.37 out of a total of 30. A better score than that out of 20.

Table 16—29 Sub-test es04 Standard error of measurement

Standard error of measurement	1.37

16.5 Conclusions

The important conclusions we are able to draw from this sub-test lie in two areas. Firstly, there are two points that center on the tasks and skills.

- We are able to add further weight to the argument in favor of using items based on skill 22.1:

  Understanding information in the text, not explicitly stated, through making inferences

  In three of the four MCQ items that focus on this skill, we have achieved good results.

- The error identification and error correction items focus on assumed knowledge, and reveal a significant lack of precision. These items would seem to be testing skills that are in need of improvement.

Secondly, we have a number of lessons-to-be-learned about the test writing process:

- Problems arise when item writing is uncoordinated
- It is of value for item writers to analyze the actual responses of candidates in order to learn how better to write items
- Stem and option writing requires a good deal of care, and is enhanced by teamwork, ensuring that unprejudiced eyes lead the revision process.

17 SUB-TEST IN03

17 SUB-TEST IN03

17.1 TEXT
17.2 TASK TYPES
17.2.1 *Four-option multiple choice questions*
17.2.1.1 Facility values
17.2.1.2 Discrimination indices
17.2.1.3 Discussion
Item #5
Item #6
Items #8, #9, and #10
Item #1
Item #7
Item #3
Item #4

THIS TEST was the Language B English counterpart of LA sub-test es04. It was written and administered to candidates taking the entry test to the University of Granada in July 1994. The sub-test was taken after a test in LA at the Faculty of Arts, University of Granada, in a single morning session, and in parallel with tests in LB German and French. The results of this test, and those produced from the LA test were used to select those who would later take a sub-test in LB listening comprehension.

The single most important factor that may have had an impact on candidate performance in this test is the fact that the circumstances in which the LA sub-test was administered were the cause of conflict. The administration of the LA sub-test was wholly unsatisfactory causing protests from many of the candidates both during and after the session. These later led to an internal disciplinary inquiry. During the session candidates encountered what they believed was an error in the paper, but they did not receive a satisfactory explanation from its author. The test paper had not been produced according to the established specification or schedule, it had not been revised or trialled. The member of the Examination board responsible for its administration and correction failed to apply the markscheme correctly, leading to inconsistencies in marking. There also appeared a number of clerical errors in many of the scripts meaning that scores were totaled incorrectly. As a consequence the Examination board decided to alter the weighting of the LA and LB tests to give the former less value. The ratio was changed from 1:1 to 2:3. All of these elements lead us to believe that candidate performance may have been negatively affected as a result.

Three hundred and sixty-two candidates took sub-test in03, and, other factors not prevailing, we were reasonably confident that the results produced would be representative, and that these would confirm, or not, the conclusions we had drawn from the more limited sample that took es04. As these candidates would also continue their studies in Granada, we knew that we would be able to make a follow-up study of their performance.

Table 17—1 Sub-test in03 "A skeleton in the cupboard" (*The European*, 3-9 June 1994; Élan:30)

Birna Helgadottir on the literary triumph of an Icelander who is also a New York yuppie prince

Olafur Olafsson

Absolution Orion, ,12.99

(1) IF OLAFUR OLAFSSON were a character in one of his own novels, critics would no doubt think him a far-fetched fantasy. The son of one of Iceland's best-known writers, Olafur Johann Sigurdsson, the young Olafsson joined Sony, rising to become, at the height of New York yuppie fever, the president of Sony Electronic Publishing - one of the youngest company presidents in US business history. At the same time, he published his first novel to rave reviews back in Iceland.

(2) Now, at 31, both Olafsson the executive and Olafsson the novelist are flourishing to an extraordinary degree. Electronic Publishing, riding the crest of the video game boom, is the fastest-growing branch of the Sony empire. Meanwhile its president, already the best-selling author in Iceland's history has had his first novel in English, Absolution, hailed by critics in the US and Britain as a worthy successor to Dostoevsky's Crime and Punishment.

(3) Olafsson has chosen as the subject of his first "international" work how embittered old age follows a wasted life. Absolution is a study of a man whose life has been poisoned by a "little crime" of treachery committed in youth. The novel takes the form of the memoirs of Peter Peterson, an Icelandic expatriate living the life of a recluse in New York's Park Avenue. He has amassed a fortune through a lifetime of dubious business practices, is cruel to his estranged family and has no faith in God or humanity. His only interests are old movies, wine and the charms of his housekeeper.

(4) We are gradually introduced to the old misanthrope's younger self, an idealistic youth, then known

by his proper Icelandic name Petur Petursson. Young Petur leaves his native Reykjavik to follow Gudrun, the girl he worships, to Copenhagen, just before the Nazi occupation. It is here, in a blind fog of jealousy and shattered illusion, that he carries out the deed that transforms him into the bitter, amoral Peterson. The narrative, which zips back and forth between past and present, pre-war Scandinavia and modern Manhattan, is sandwiched between the musings of its "editor", a sensitive young Icelander who has discovered Peterson's manuscript after his death.

(5) In Iceland the novel has sold 13,000 copies, one for every 20 inhabitants. Olafsson's success there is in many ways due to the fact that he embodies the curious paradox of the Icelandic psyche.

(6) On the one hand, it is a materialistic nation with one of the highest living standards in the world. In recent years the traditional devotion to the Protestant work ethic has often combined with growing consumerism to produce the worst kind of yuppie excesses.

(7) At the same time, Iceland is probably the world's most literate and literary nation. This is the country that produced not just the medieval sagas - Europe's first vernacular prose - but other more esoteric literary efforts. The world's first Basque dictionary was written by a farmer in North-West Iceland - he had learned the language from Basque fishermen. The only epic poem ever written in honour of the medieval Balkan hero Skanderbeg was composed by an otherwise unknown Icelandic clergyman in 1861.

(8) Literacy rates have been nearly 100 per cent since at least the 18th century. With 400-500 published here every year, Iceland produces more books per head of population than any other country. One in ten Icelanders will be published in his or her lifetime - writing a book is seen almost a form of national service.

(9) There is a poetic, mystical side to the Icelanders - most claim to believe in God and most believe in elves and ghosts, Poetic ability is considered the hallmark of a noble personality, and artists and writers have a special

dispensation, in a nation where hard work is considered beneficial for its own sake, to lead a ramshackle, layabout lifestyle. As the Icelandic Nobel prizewinner Laxness put it, "since time immemorial, the Icelandic nation has had to battle with men who call themselves poets and refuse to work for a living."

(10) But Olafsson is a poet who works hard for his living. He has also fulfilled another important Icelandic dream - to become famous abroad: he has been featured in Fortune magazine as a high-flying company president and the New York Times Book Review as a promising novelist.

17.1 Text

The text (*The European*, 3-9 June 1994; Élan:30) was edited down to 730 words and the version presented was in word-processed format. Numbers were added to identify the paragraphs, and some of these paragraphs were identified in the question stems. This procedure was adopted to counter the extra length of the text by comparison with the earlier sub-tests. Adding paragraph numbers to the item stems speeds up the reading required to locate the answer.

The topic is a book review, requiring no specialist knowledge, and believed unlikely to favour any particular group of candidates. Both the author and the book were thought to be sufficiently remote from the candidates' context that it was highly improbable anyone would have heard of either.

17.2 Task types

As in sub-test es04, three task types were used, four-option MCQs, error identification, and error correction items.

17.2.1 *FOUR-OPTION MULTIPLE CHOICE QUESTIONS*

17.2.1.1 *Facility values*

The overall performance of these items was disappointing. Of ten MCQ items only two — #5 and #6 — produced FV scores within out target range, and the first of these was defective, as we will see shortly. These two intermediate FV items aside, the remainder were either easy, or very easy.

Table 17—2 Sub-test in03 Items #1-10 Target skills (Munby 1978:126-131)

Items	*Target skills*
#1	Interpreting text by going outside it, "reading between the lines" (Skill 34.2)
#2, #8, #9, #10	Understanding explicitly stated information (Skill 20) Understanding information in the text, not explicitly stated, through making inferences (Skill 22.1)
#3	Understanding the communicative value (function) of sentences and utterances with explicit indicators (Skill 26.1)
#4, #6	Understanding explicitly stated information (Skill 20)
#5	Deducing the meaning and use of unfamiliar lexical items, through contextual clues (Skill 19.2)
#7	Selective extraction of relevant points from a text, involving the coordination of related information (Skill 41.1)

17.2.1.2 *Discrimination indices*

In contrast, the DI scores achieved by the same set of items were much more promising. None were negative, and only two items achieved a poor level of discrimination, and one of these was #5, which we have already mentioned was defective. Of the remainder, four can be classified as average discriminators, coming below our target. The other items were all very good discriminators, ranging in scores from 0.41 to 0.52.

17.2.1.3 *Discussion*

In this section we will look at the items in groups according to their performance and the skills they were designed to test. We will start by discussing the two items that scored the best Facility values.

ITEM #5

We will deal with this item first because it is defective, and clearly so, and because it is a good example of the difficulties of producing good MCQ items without trialling.

Figure 17—1 Sub-test in03 Items #1-10 Facility values

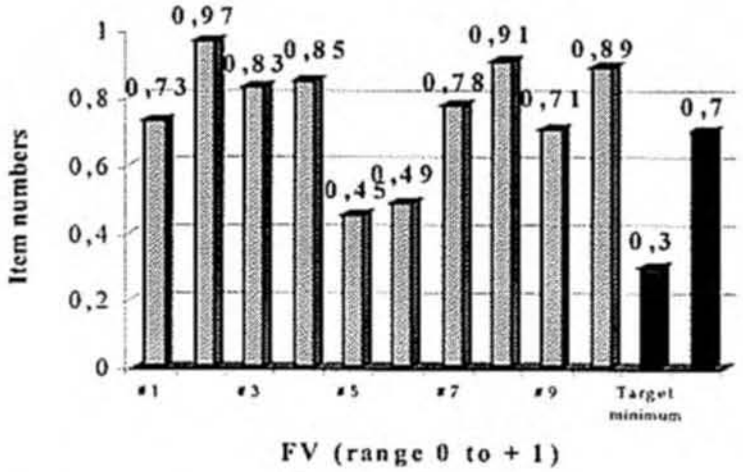
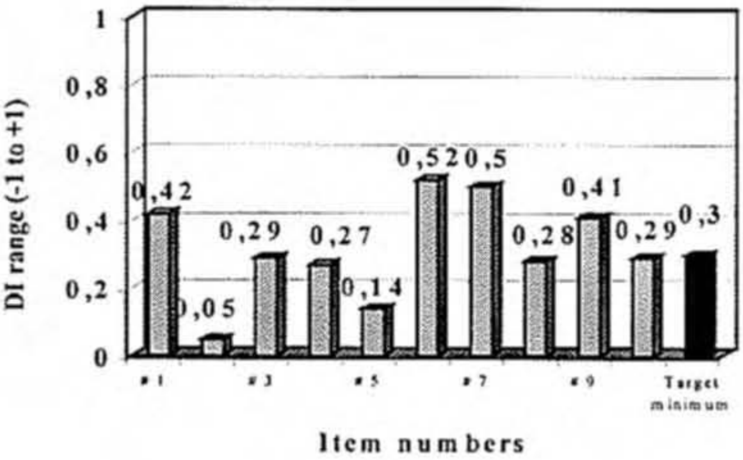


Figure 17—2 Sub-test in03 Items#1-10 Discrimination indices



This item, and all of the test, was revised by members of the Examination board prior to administering the test, but at no point were the options offered questioned. The text of the item, shows that the key response was intended to act as a synonym for the target word *estranged*. There are two defects in the item. The first is apparent without reference to the text, as none of the options can be seen to be a close synonym of the target word. The second reason can only be perceived by referring to Paragraph 3 of the text.

Table 17—3 Sub-test in03 Item #5 Question stem and options (✓ indicates key response)

5 In the phrase "... to his estranged family ..." (Paragraph 3), estranged means the same as ...

A. unfriendly. ✓

B. foreign.

C. faraway.

D. curious.

The sentences from which candidates are intended to deduce the meaning of *estranged* by using contextual clues are:

> *The novel takes the form of the memoirs of Peter Peterson, an Icelandic expatriate living the life of a recluse in New York's Park Avenue. He has amassed a fortune through a lifetime of dubious business practices, is cruel to his estranged family and has no faith in God or humanity.*

In this context, the clearest meaning of *estranged* is #5C *faraway*. This conclusion was reached by nearly half of the candidates, whereas only 15% chose the key response, making this the least popular of all of the alternatives. As we have said in the context of other items, a well-written set of options should produce similar FVs for each of the incorrect alternatives. In this case we have a badly-written item, which does actually produce a good spread of responses.

Table 17—4 Sub-test in03 Item #5 Facility values of all options

Item	*#5A*	*#5B*	*#5C*	*#5D*
Actual responses	56	60	164	67
FV	0.15	0.17	0.45	0.19

In the context of the actual administration of the test, the item was counted as valid, and option #5A was scored as correct, maintaining the key response as predicted. This was felt to be the fairest way of managing the problem without reducing the number of valid items, and in the light of the performance of the Error identification and Error correction items which we discuss below. This was the option that produced the best discrimination index score. Option #5C, the most popular of the responses, and that which might be considered the correct measure of the target skill being tested, actually provided us with a negative DI.

Table 17—5 Sub-test in03 Item #5 Discrimination indices of all options (shading indicates the key response)

Item	*#5A*	*#5B*	*#5C*	*#5D*
DI	0.14	0.01	(0.10)	(0.05)

All in all, we can say that the item has provided much food for thought from a technical point of view, and little of value for the immediate question of assessing candidates' ability to use their reading skills.

ITEM #6

"Understanding explicitly stated information" is the target skill on which this item focuses. This is one of the more straightforward skills, and so it would not be predicted to produce a low FV. However, the fact is that the FV for this item was 0.49, making it stand out in this set of Facility values. The key response to this question, #6D, is based on information found in Paragraph 4, and the difficulty may lie in the fact that the location of the answer is not specified in the stem by reference to the paragraph number. Candidates may have found that any of the alternatives, with the exception of #6C, may be correct. In this case it is a case of eliminating options which do not appear in the text and/or associating the phrase

> *The narrative, which zips back and forth between past and present, pre-war Scandinavia and modern Manhattan...*

with option #6D.

Table 17—6 Sub-test in03 Item #6 Question stem and options (✓ indicates key response)

6 In Absolution, the narrative ...

A. Displays characteristics typical of Icelandic literature. ____

B. is a chronological account of the protagonist's life. ____

C. has been written by an "editor". ____

D. Changes perspective frequently. ✓

The FV scores for all options show that #6B proved a very strong distractor, and that neither #6A nor #6C served much purpose. Having said that, it is therefore interesting to note that this was the strongest of all of the items in terms of discrimination. The DI score of 0.52 makes it a very good item.

Table 17—7 Sub-test in03 Item #6 Facility values of all options

Item	*#6A*	*#6B*	*#6C*	*#6D*
Actual responses	10	124	33	179
FV	0.03	0.34	0.09	0.49

Why option #6B should be such a strong distractor is difficult to explain. It contradicts the key response, and the information that indicates that it is incorrect seems fairly clear in the same phrase taken from Paragraph 4 and which we have quoted above. The reason may lie in candidates' unfamiliarity with the word "zips", and their inability to deduce its meaning from the context.

ITEMS #8, #9, AND #10

These three items focused on the target skills of understanding information both explicit and implicit in the text. As could be predicted, those that were easier — #8 and #10 — failed to reach the target minimum for discrimination, but all of them functioned reasonably well.

The key response to item #8 can be understood by a careful reading of the third sentence of Paragraph 8. It restates the proposition of the stem and option #8B, and the difficulty lies in candidates' ability to interpret the phrasing of the text:

> *One in ten Icelanders will be published in his or her lifetime - writing a book is seen almost a form of national service.*

Table 17—8 Sub-test in03 Items #8, #9, and #10 Question stem and options (✓ indicates key response)

8	One out of ten Icelanders ...	
A.	has bought a copy of Absolution but has not read it.	
B.	has published a book.	✓
C.	has read Absolution.	
D.	is a poet.	
9	In Iceland people who think themselves artists or writers ...	
A.	lead bohemian lives.	
B.	are social outcasts.	
C.	refuse to work for a living.	✓
D.	believe in elves and ghosts.	
10	Olafsson has fulfilled an important Icelandic dream because he...	
A.	embodies the traditional image of the writer.	
B.	does not believe in elves and ghosts.	
C.	has become famous abroad.	✓
D.	works hard.	

It is important for candidates here to be able to interpret the modal "will" as indicating a generalization that is echoed in the present perfect form of the option.

In item #9, the statements which are embodied in the stem and each of the options could all be correct according to the manner in which the text is read. The key response is found in Paragraph 9, but candidates need to interpret several phrases in order to reach the conclusion that the item presents.

> *...artists and writers have a special dispensation, in a nation where hard work is considered beneficial for its own sake, to lead a ramshackle, layabout lifestyle.*

The item serves as a summary sentence for the entire phrase, reproduced above.

Item #10 is surprising in that the stem and key response echo two phrases that appear in the text:

> *...another important Icelandic dream - to become famous abroad...*

It seems too easy, in fact, but the item has functioned surprisingly well.

In all three items, the range of responses attracted by the distractors has been wide, showing that the incorrect options have functioned poorly despite the good results achieved in terms of discrimination. We can conclude, therefore, that technically these items leave something to be desired, but in practical terms they serve a purpose.

ITEM #1

Another example of a poorly written item that has technically functioned well is item #1. The target skill involved in this case was that of "reading between the lines", and the key option is distant from the content of the text.

Table 17—9 Sub-test in03 Item #1 Question stem and options (✓ indicates key response)

1	The growth of the electronic publishing industry ...	
A.	continues to amaze the business world.	✓
B.	has failed to live up to expectations.	
C.	appears to have reached a plateau.	
D.	has past its peak.	

However, the key response is the only option that is positive, indicating that it is clearly correct, or that it is incorrect and that there are only three valid alternatives. Candidates need not read the text in order to come to this conclusion. However, once they read the text and find all of the positive information relating to the electronic publishing industry, particularly in Paragraph 2, they can have little doubt about it. The FV was only just above our target, and the DI score was 0.42, showing that the item discriminated well. Option #1C was the strongest of the three distractors, but the range was wide: FV 0.02 to 0.17, indicating that they functioned poorly.

ITEM #7

The target skill on which this item focuses is one that we have not tested in previous sub-tests.

Selective extraction of relevant points from a text, involving the coordination of related information (Munby 1978:126-131 Skill 41.1

Although the item stem refers to Paragraph 5, this is just the location of the target phrase and the options echo or quote elements that appear in Paragraphs 6, 7, and 8.

Table 17-10 Sub-test in03 Item #7 Question stem and options (✓ indicates key repsonse)

1 "... the curious paradox of the Icelandic psyche ... (Paragraph 5) is that as a nation Icelanders ...

A. enjoy high living standards but act like yuppies. ____

B. are materialistic and yet highly literate and literary. ✓

C. are devout Protestants who have an affinity for the Basques. ____

D. publish huge numbers of books each year though they read few. ____

The key response is a composite of information that appears in both Paragraph 6 and Paragraph 7. Candidates might be tempted to jump at #7B because the first element of this option is the first response they would find in a linear reading of the text. However, the fact that the other options are strong should entail a deep reading of the paragraphs concerned and some thought in order to put the response together.

The item was not as difficult as we would have liked, and the distractors performed poorly, with a range of FV scores from 0.01 to 0.14. Nonetheless, item #7 did discriminate very well with a DI of 0.50.

ITEM #3

This item focuses on an element of written production, which candidates may or may not be familiar with, in order to test their ability to use the skill of

> *Understanding the communicative value (function) of sentences and utterances with explicit indicators (Munby 1978:126-131 Skill 26.1)*

The item stem focuses on the phrase *a "little crime" of treachery* and asks candidates to select the appropriate interpretation of the communicative value of the inverted commas.

Table 17—11 Sub-test in03 Item #3 Question stem and options (✓ indicates key response)

3 When the writer uses the words "little crime" between inverted commas she does so because she ...

A. is using a recognized literary term. ____

B. wants to produce an ironic effect. ✓

C. intends it as reported speech. ____

D. is quoting from the text. ____

In order to do so candidates must move beyond the text itself and apply their intertextual knowledge. In practice, the supratextual convention is not difficult to identify, and the FV score for the item demonstrates its ease. However, the DI score shows it is an average discriminator at DI 0.29, and it almost enters our target range. The FV scores for the alternatives are poor, though, with a wide range. We would have hoped that each option might attract something like 20 actual responses, but only #3D was chosen by a significant number of candidates, and this was intended to be a strong distractor.

Table 17—12 Sub-test in03 Item #3 Facility values of all options

Item	*#3A*	*#3B*	*#3C*	*#3D*
Actual responses	4	299	14	43
FV	0.01	0.83	0.04	0.12

ITEM #4

We have already seen that item #6 performed well. Item #4 was focused on the same skill, but was much easier. It did, however, produce an average DI of 0.27.

The item itself offers a range of options, all of which superficially could be correct. In order to reach the key response, #4C, it is necessary to identify #4B as incorrect, because these are mutually exclusive. Option #4A is the most obviously incorrect, as candidates discover fairly early on in their reading of the text. And option #4D is incorrect, but is an echo of a phrase that does appear in Paragraph 2 and distracts well.

Table 17—13 Sub-test in03 Item #4 Question stem and options (✓ indicates key response)

4 Absolution ...
A. deals with an Icelandic theme. ____
B. was first published in Icelandic. ____
C. is Olafsson's first novel in English. ✓
D. made Olafsson the best-selling author in Iceland's history. ____

The FV scores for all of the options show how only one candidate failed to discount #4B, and that of the other two distractors one was stronger.

Table 17—14 Sub-test in03 Item #4 Facility values of all options

Item	*#4A*	*#4B*	*#4C*	*#4D*
Actual responses	14	1	309	39
FV	0.04	0.00	0.85	0.11

From the point of view of item writing, this item does indicate that options that in one way or another echo the passage can serve as useful distractors.

17.2.2 *Error identification and Error correction*

These task types were new to candidates, and as they are not frequently used in EFL teaching, we believed that some candidates might not perform as well as they should. In order to reduce this to a minimum we used a highly explicit rubric, which listed the type of error and the number of errors of each type, and we included an example in the task (Table 17—16).

These items were based on a selection of target skills drawn from Munby (1978) (Table 17—15). We will now begin to look at the items in terms of the overall pattern of Facility values and Discrimination indices they produced.

17.2.2.1 *Facility values*

The range of Facility values produced by these items was both narrow and low: 0 to 0.33. Only three error identification, and two error correction items reached our target minimum. The remaining items fell below this level. This means that all of the items in the two sets were difficult or very difficult for candidates. This is exactly the opposite of the pattern that emerged with the MCQ items based on the same text.

Table 17—15 Sub-test in03 Items #11-30 Error identification and error correction target skills (Munby 1978:126-131. Shading indicates items that are not derived from this source)

Items	*Target skills*
#11-12, #19-20, and #27-28	Recognizing the script of a language: following grapheme sequences (Spelling system) (Skill 17.2), and Manipulating the script of a language: catenating grapheme sequences (Spelling system) (Skill 18.2)
#13-14	Knowledge of vocabulary
#15-16, and #29-30	Understanding relations within the sentence (Skill 28), and Expressing relations within the sentence (Skill 29)
#17-18 and #25-26	Recognizing the script of a language: understanding punctuation (Skill 17.3) and Manipulating the script of a language: using punctuation (Skill 18.3)
#21-22	Recognizing indicators in discourse for concluding an idea (Skill 35.4) and Using indicators in discourse for concluding an idea (Skill 36.4)
#23-24	Recognizing indicators in discourse for explanation or clarification of point already made (Skill 35.6), and Using indicators in discourse for explanation or clarification of point already made (Skill 36.6)

Table 17—16 Sub-test in03 Error identification and error correction items (Key responses appear in the right-hand column in bold italics. Item numbers have been added to facilitate the discussion.)

The text continues here, but 10 errors - of grammar (2), cohesion (1), spelling (2), punctuation (3), and vocabulary (2) - have been introduced. Read the text carefully and underline the errors you find. On the line provided in the second column write the correct form. One answer has been given for you as an example.

(11)~~Nontheless~~ there are signs that the Reykjavik intelligentsia is whipping up something of a (13) ~~backslap~~ against the golden boy - some of its members have slated the book, calling it formulaic, catchpenny literature, and have complained that (15) ~~they're~~ other (17) ~~icelandic~~ writers more deserving of an international following.

As a novelist Olafsson has his (19) ~~floors~~. His stately style may not be to everyone's taste, and he lacks (21) ~~lyricism atmosphere~~ is another (Example) ~~week~~ point - his anachronistic portrayal of pre-war Reykjavik cannot compare with the vivid way writers (23) ~~such like~~ Einar Karason and Petur Gunnarsson have conjured up the post-war period.

He also resorts to obvious ploys to whet (25)~~the readers appetite~~. But overall the book works extremely well both as a (27) ~~psycological~~ thriller and as a character study.

Though *Absolution* is set for publication in Germany, France and Norway this autumn, there is little chance of Olafsson giving up his day job. (29) ~~He is saying~~ he does not want to have to write to pay.

(12) *Nonetheless* there are signs that the Reykjavik intelligentsia is whipping up something of a (14) *backlash* against the golden boy - some of its members have slated the book, calling it formulaic, catchpenny literature, and have complained that (16) *there are* other (18) *Icelandic* writers more deserving of an international following.

As a novelist Olafsson has his (20) *flaws*. His stately style may not be to everyone's taste, and (22) *he lacks lyricism. Atmosphere* is another (Example) *weak* point - his anachronistic portrayal of pre-war Reykjavik cannot compare with the vivid way writers (24) *such as* Einar Karason and Petur Gunnarsson have conjured up the post-war period.

He also resorts to obvious ploys to whet (26) *the reader's appetite*. But overall the book works extremely well both as a (28) *psychological* thriller and as a character study.

Though *Absolution* is set for publication in Germany, France and Norway this autumn, there is little chance of Olafsson giving up his day job. (30) *He has said/says* he does not want to have to write to pay.

Figure 17—3 Sub-test in03 Error identification items Facility values

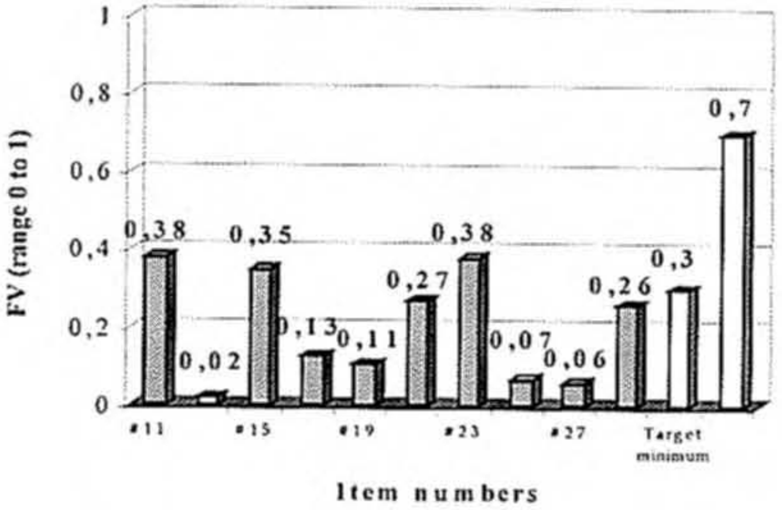


Figure 17—4 Sub-test in03 Error correction items Facility values

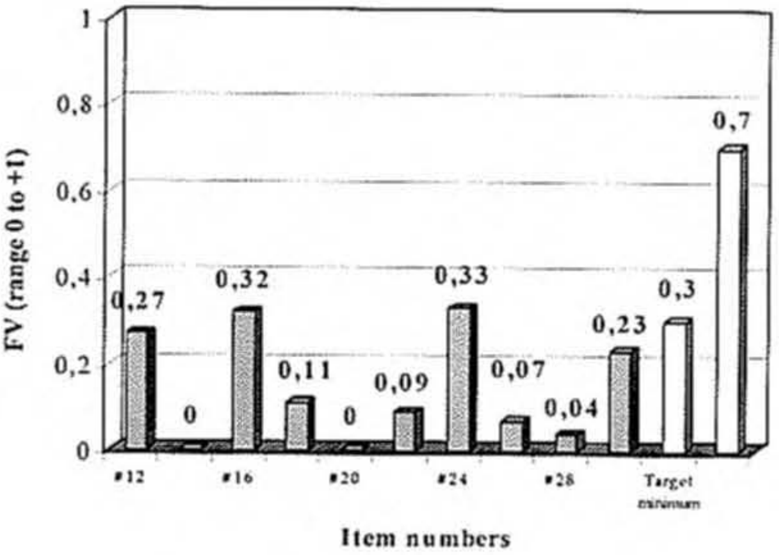


Figure 17—5 Sub-test in03 Error identification items Discrimination indices

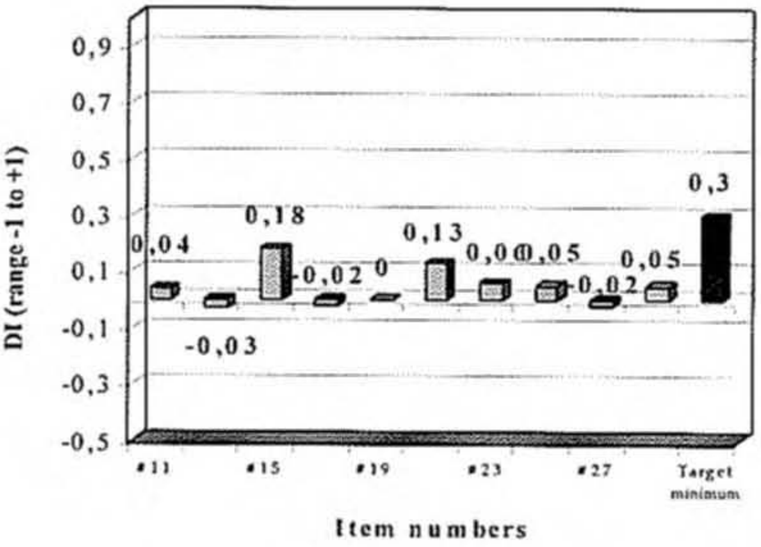
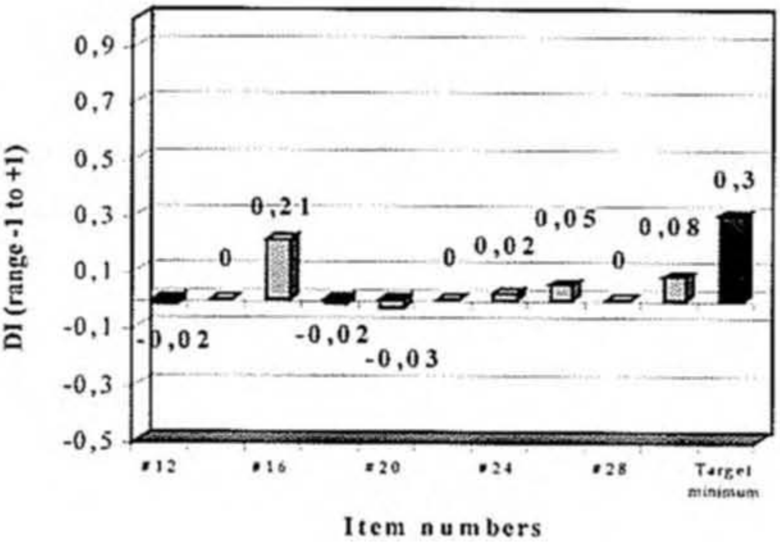


Figure 17—6 Sub-test in03 Error correction items Discrimination indices



17.2.2.2 *Discrimination indices*

In general, low FV scores mean low DI scores, and this sub-test was no exception. None of the items reached the target minimum, and six produced negative DI values. The highest discrimination index score was for item #13, which reached 0.21, an average level of discrimination.

17.2.2.3 *Discussion*

In global terms, we can say that this set of items has proved far too demanding of candidates. But we must ask ourselves whether this is a consequence of the inherent difficulty of the target skills involved, a function of the specific task type, or of the individual items included in the paper. If the construct validity is proven, then the scores represent an accurate judgement on the capabilities of the candidates; if either the task type or the individual items are called into question, then the results have more to do with poor test design that actual candidate performance.

We will now group the items according to the target skills involved and discuss their performance in more detail.

Items #11, #19, and #27 all focus on spelling in one way or another. The first was what we considered a straightforward error of omission — *nontheless* for *nonetheless* — the second a more complex error involving the use of a homophone, and the third another relatively simple error of omission — *psycological* for *psychological*. It was our belief that both #11 and #27 would score an FV within our target range, and that these were likely to be towards to top of that range.

We thought that #19 was likely to be comparatively more difficult, but that it too would score an FV comfortably within the range. Item #19 focussed on the recognition of a homophone, and the necessary correction of the written form. This exercise in precision was included as being a test of candidates' ability to identify a highly specific error of this kind. The error in question —*floor* or *flaw* — was not "authentic", in that it was not an error that had been encountered in student work. It was, however, felt to be a plausible error that might occur, and clearly was an error that the text enabled us to use.

The FV scores for these items indicate that our predictions in terms of difficulty were far from correct. The three items scored FVs of 0.38, 0.11, and 0.06, respectively. This means that only the first achieved our target minimum, and scored the same FV as item #22, both of which were the highest in the whole set.

The error correction items based on these items were naturally enough, lower scoring still. Item #12 scored an FV of 0.27, which makes it difficult, but it was close to our target; but #20 scored

0.06, and #28 only 0.04, meaning that virtually nobody was able to correct them successfully.

What is more disappointing is that the DI values achieved by items #11 and #12, which apparently performed better than the others, were actually very low for error identification, and negative for correction. Items #19-20 and #27-28 also produced negative or zero DI scores.

Item #13, based on the knowledge of vocabulary was very difficult, with an FV of 0.02, and the corresponding error correction item #14 scored 0, i.e. nobody was able to respond correctly. This item was probably ill-conceived in that the words involved — *backslap* and *backlash* — are probably too obscure in meaning and too close in morphology, for candidates to be aware of the difference. The DI value for error identification was actually slightly negative, meaning that it was not only difficult for candidates to find the error, but that those candidates who did find it were actually the weaker candidates overall. Correction did not discriminate because no one was able to correct it.

The two items that focused on "relations within the sentence", #15 and #29 achieved better FVs: 0.35 and 0.26, respectively. What's more, the FVs for correction were also comparatively better 0.32, and 0.23. The differences between these are marginal, meaning that those who were able to locate the errors were usually able to correct them. The first of these items entailed find the incoherence of the phrase

> *...and have complained that they're other Icelandic writers more deserving of an international following.*

A careful reading would lead candidates to the realization that with no relative pronoun they would need to make some kind of correction, and substitute *there are* for *they're*. In the case of item #29 the task was apparently almost as easy. Candidates needed to identify the inappropriate verb form with which the sentence begins:

> *He is saying-he does not want to have to write to pay.*

This involves relating the sentence to the rest of the text and understanding the communicative function it fulfils. While identification of the error was reasonably good under the circumstances, it is surprising that very few candidates were able to correct it adequately, and that neither the identification of the error nor its correction served to discriminate between candidates.

Similarly, it was disappointing to discover was that the two items that focussed on punctuation, #17 and #25, were extremely difficult for candidates, although the majority of those who were able to identify the error succeeded in correcting it. The first of these simply required candidates to spot the lower case *i* used with *icelandic*, which is acceptable in Spanish, but not so in English. The second, which we also considered straightforward, consisted of identifying the missing apostrophe of a possessive form — *to whet the readers appetite* — and to correct the phrase. This latter item was found by very few, and was corrected by only a few of these. It did not discriminate between candidates.

The two remaining items were thought to be the most difficult on the paper as each involved recognizing and using indicators in discourse. In fact, these produced good facility values, given the overall nature of performance, except that very few of the candidates who were able to identify the error in #21 were subsequently able to correct it.

Item #21 was based on the lack of a full stop and capital letter, meaning that two sentences ran into one:

> *...and he lacks lyricism atmosphere is another weak point...*

Many candidates completely failed to understand the error, and would appear to have read *lyricism atmosphere* as a compound noun. Those who were able to identify it attempted different means of correcting, many of which were acceptable even though they were different to the key response. In these cases, the item was marked correct if, in the context of the passage, the correction proposed was considered acceptable.

The use of a connective phrase — *such as* — was the basis of the remaining item. In this case, the candidates who identified the item were generally able to correct it, but this item, too, failed to discriminate.

Perhaps the most striking conclusions that can be made from these sets of items are in two directions: firstly, they have to do with the perception on the part of the test writer of the types of error candidates should/would be able to identify. Secondly, they concern the levels of ability demonstrated by candidates in the MCQ items, and the comparable levels in these tasks.

The skills on which the error identification and error correction items were based do not appear to be of any tremendous difficulty. Rather, they are on the whole somewhat "tame" items. Colleagues in the Examination board and in the Department did not consider these items unduly difficult, nor did they believe that they lacked content or construct validity. In some quarters the conclusion reached was simply that the standard of English attained in secondary schooling was "that poor". However, it seems highly unlikely that such a simplistic explanation should be correct. We feel that the reasons behind these results probably lie in a combination of these factors and, perhaps, others that we have yet to identify.

One thing is clear, by carrying out a comparison between the two parts of the test, we have identified a strange discrepancy. The correlation between candidates scores for the two weighted elements of the sub-test, the MCQ items out of a total of 10 and the Error identification and Error correction items out of a total of 20, was as low as 0.12. While we would expect there to be a difference between the level of passive skills demonstrated through the MCQ items, and active skills demonstrated through the Error identification and Error correction items, we would not predict such a great difference. A correlation of only 0.12 indicates virtually no relationship between the two, and that is almost impossible to believe. This finding alone must incline us to the conclusion that the greater part of the responsibility for these results lies in elements other than the candidates' level of English.

17.3 Descriptive statist ics

Even the most superficial analysis of the four histograms that depict aspects of the results of this sub-test reveals the discrepancy between the scores achieved by candidates in the MCQ items and in the Error identification and Error correction items. The curve of is strongly negative with skewness recorded at -0.78. This influences the weighted scores, but to a lesser extent, producing a figure of –0.24. The histogram representing the scores out of 20 for the Error identification and Error correction items is totally different (). Here the skew is almost as great, but it is positive at 0.70.

17.4 Reliability

Despite all of the difficulties that we have encountered in the individual items and particularly in the manner in which candidates have responded to the task type, we find that the average reliability coefficient for the test is quite good. In order to calculate this we have viewed the test as a 30 item whole, rather than taking into account the weighting in any way, as the construction of the split halves does not permit this. The data presented in Tables 17—21, 17—22, and 17—23, therefore, reflect the statistics presented and present a more "normal" distribution.

If we were to produce a revised version of this same test, by adding nine items of the same nature to a 30 item, unweighted sub-test, we would be able to attain a more than acceptable level of reliability.

In this instance, due to the nature of the formulae, we are only able to offer the S_e based on the unweighted total out of 30. This produces a figure of 1.99, meaning that we can be 68% certain

Figure 17—7 Sub-test in03 Frequency distribution of scores for MCQ items

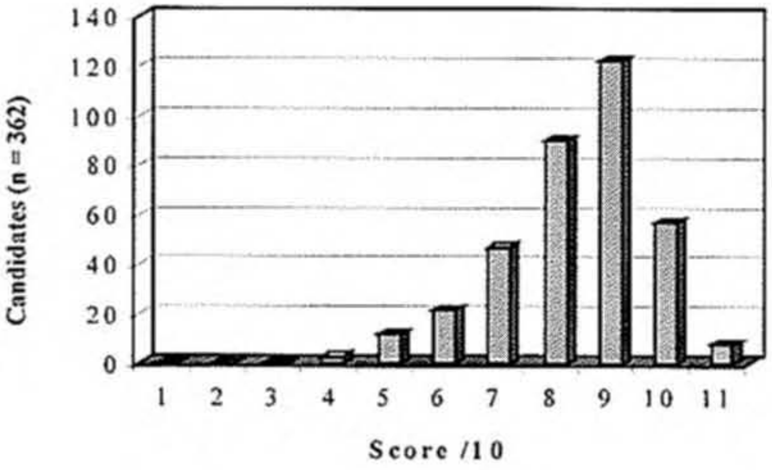


Figure 17—8 Sub-test in03 Frequency distribution of scores for error identification and error correction items

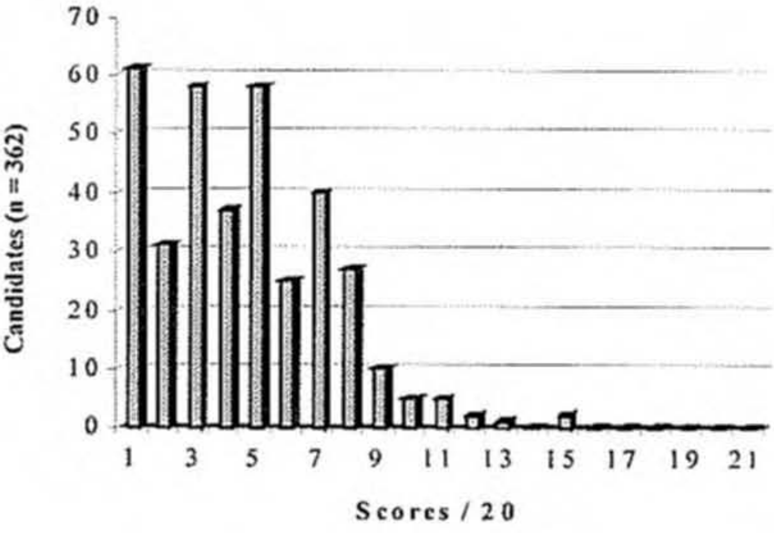


Figure 17—9 Sub-test in03 Frequency distribution of weighted total scores

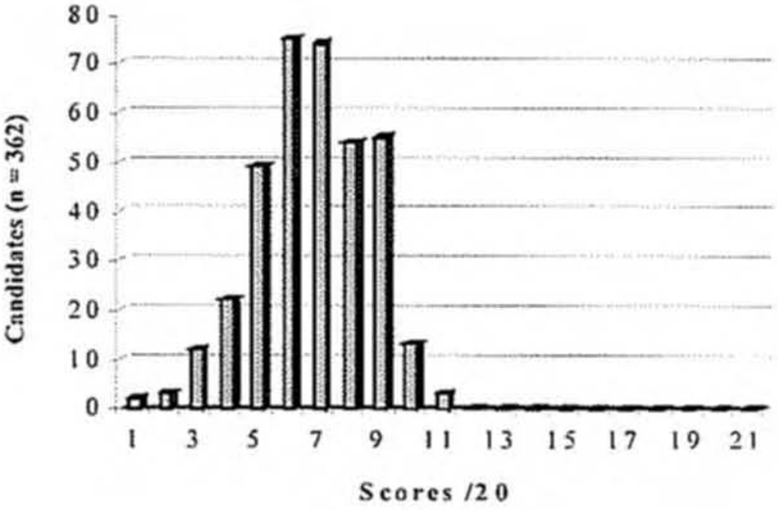
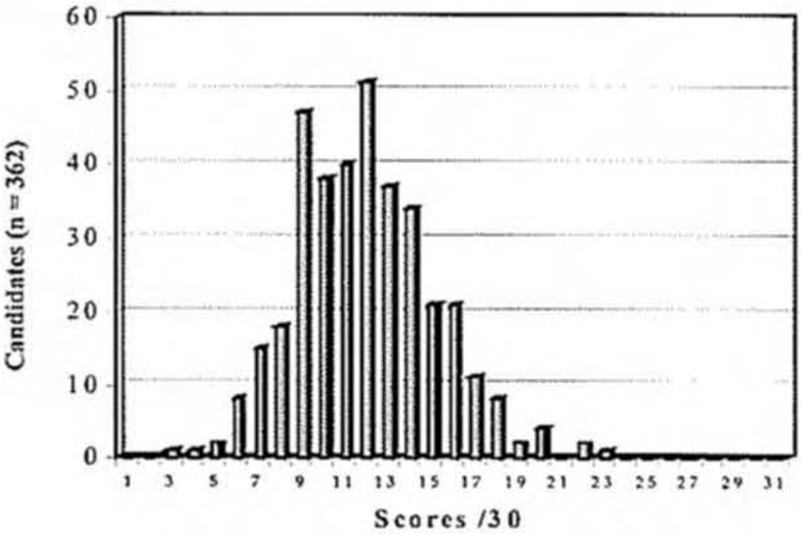


Figure 17—10 Sub-test in03 Frequency distribution of unweighted total scores



any candidates' true score would fall within a range of their observed score ±1.99. While this is greater than the S_e for sub-test es04, it is still an improvement on the margins we found in the 20 item sub-tests taken by Cohort A.

Table 17—17 Sub-test in03 Descriptive statistics for MCQ items (score out of 10)

Mean	7.32
Median	8.00
Mode	8.00
Standard error	0.07
Standard deviation	1.39
Skewness	(0.78)
Level of confidence: 95%	0.19

Table 17—18 Sub-test in03 Descriptive statistics for error identification and error correction items (score out of 20)

Mean	3.53
Median	3.00
Mode	0
Standard error	0.14
Standard deviation	2.74
Skewness	0.70
Level of confidence : 95%	0.28

Table 17—19 Sub-test in03 Descriptive statistics for weighted scores (out of 20) Mean 5.55

Median	5.50
Mode	5.00
Standard error	0.09
Standard deviation	1.83
Skewness	(0.24)
Level of confidence : 95%	0.19

Table 17—20 Sub-test in03 Descriptive statistics for unweighted total scores (out of 30)

Mean	10.86
Median	11.00
Mode	11.00
Standard error	0.17
Standard deviation	3.22
Skewness	0.39
Level of confidence : 95%	0.33

Table 17—21 Sub-test in03 Average of split-half coefficients of reliability

Overall average	0.619

Table 17—22 Sub-test in03 Reliability of a lengthened version

Spearman-Brown Prophecy Formula (39 items)	0.991

Table 17—23 Sub-test in03 Standard error of measurement

S_e	1.99

17.5 Conclusions

The conclusions we reach from our analysis of this test, build on those for sub-test cs04. Again, we have found that the longer texts, and the increased number of items improves the overall performance of the test. We have also interpreted candidate responses and found that skills appear to operate in combinations, even though we have attempted to target our items on individual skills. Further, we have seem the effectiveness of items that aim to test skill 20 — *understanding explicitly stated information.*

18 MEETING CRITERIA: THE STRENGTHS AND NATURE OF RELATIONSHIPS BETWEEN MEASURES

18 MEETING CRITERIA: THE STRENGTHS AND NATURE OF RELATIONSHIPS BETWEEN MEASURES

18.1 ARE THE SUB-TESTS WE DESIGN AND USE IN THIS STUDY VALID?
18.1.1 *Response validity (condition 1)*
18.1.1.1 Conclusions
18.1.2 *Predictive validity (condition 6)*
18.1.2.1 Cohort A
Discussion
Conclusions
18.1.2.2 Cohort B
Discussion
18.1.2.3 Conclusions
18.1.3 *Concurrent validity*
18.1.3.1 Cohort A
University entrance examination and Secondary school average
University entrance examination and Language A reading skills
University entrance examination and Language B reading skills
University entrance examination and the OPT
Secondary school average and all other measures
Language B reading skills and the OPT
18.1.3.2 Cohort B
University entrance examination and Secondary school average
University entrance examination and Language B reading skills
University entrance examination and the OPT
Secondary school average and all other measures
Language B reading skills and the OPT
18.1.3.3 Conclusions
18.1.4 *Construct validity*
18.1.5 *Criterion validity*
18.1.5.1 Conclusions
18.1.6 *Content validity*
18.1.6.1 Conclusions
18.1.7 *Face validity*
18.2 ARE THE SUB-TESTS WE DESIGN AND USE IN THIS STUDY RELIABLE?
18.3 DOES THE DATA GENERATED IN OUR RESEARCH LEAD TO ANY OTHER SIGNIFICANT CONCLUSIONS?
18.3.1 *Can reading skills be used as an efficient measure in our real-world situation?*
18.3.2 *Conclusions*
18.4 ITEM ANALYSIS
18.4.1 *Matching exercises*
18.4.2 *True/False/Don't know*
18.4.3 *Multiple choice questions*
18.4.4 *Error identification and Error correction*
18.4.5 *Target language skills*
18.4.6 *Item writing*
18.4.7 *Facility values and Discrimination indices*

THE PURPOSE of this chapter is to analyze the results of our research against the criteria and performance targets established during the design phase (Part II). We divide the chapter into several parts. First, we deal with the questions of validity relevant to the longitudinal study, which was our principal objective. Then we look at the validity of measures involved in the cross-sectional study, and we continue with by analyzing the reliability of the tests. After this, we take into account the measures that demonstrate the practicality of our test instruments and analyze their performance in terms of the principal indicators of item analysis, namely the Facility values and Discrimination indices, and the target skills tested. Finally, we consider the test design process and discuss the combinations of task types and target language skills that have performed well, and that we believe could be exploited effectively in further tests.

In Table 18—1 we reproduce the criteria that we decided to use in the longitudinal part of our study and, in the later sections we present those we employed in the cross-sectional part. These are expressed as correlation coefficients derived from the instruments we have used.

Table 18—1 Summary of test implementation criteria, means of testing, and targets relevant to our longitudinal study

Sub-test performance criterion	*Means of testing*	*Target parameters*
Response validity (Carroll's condition 1)	Standard error of measurement	<1
Predictive validity (Carroll's condition 6)	Correlation coefficients between LA and LB sub-tests and General translation A-B and B-A	≥0.65 at 5% significance

In our discussion of the results we aim

- To draw conclusions about the strength of the correlation coefficients produced.
- To explain, as far as we are able, the discrepancies between our results and our targets.

- To answer the general and specific research questions on which our study is based.

18.1 Are the sub-tests we design and use in this study valid?

This is one part of our basic research question. In Chapter 7, we have already defined the different types of validity contemplated in Classical Test Theory. Here we will answer the question with reference to two of the validities on which our study is based: response validity and predictive validity. We do so in order to fulfil the two essential conditions established by J.B. Carroll (1981) in his research design model, which we have adopted for our study.

18.1.1 *RESPONSE VALIDITY (CO NDITION 1)*

We have described in Chapter 8, how the STANDARD ERROR OF MEASUREMENT (S_e) is the means by which we establish indirectly whether Carroll's condition 1 has been met by our tests. Crocker and Algina (1986:146-152) emphasize the need to ensure the S_e of any test is as low as possible and we have set our target for sub-test performance at $S_e<1$. Table 18—2 shows the actual performance of the sub-tests involved in the major part of our study.

Table 18—2 Standard error of measurement (S_e) calculations for four of our sub-tests

Cohort	*Sub-test*	S_e	*Unweighted total*
A	es01.2	1.62	20
A	es02	1.74	20
A	in02	1.95	20
B	in03	1.99	30

We can interpret these values as meaning that we are 68% certain that the true score of our candidates lies within a range of the observed score plus or minus the S_e. We can be 95% certain that it lies within a range of the observed score plus or minus $2S_e$. With total scores out of twenty, in the case of the first three tests, and out of thirty for the fourth, we must be clear that these are high margins of error. Although they are comparable to those described in the literature (Crocker and Algina 1986:151) they are far from satisfactory. In the administration of the sub-tests, and their use to make decisions on entry to the degree program, these wide margins of error will almost certainly have led to the unfair exclusion/inclusion of some candidates.

18.1.1.1 *Conclusions*

In part response to our specific research question

- **Are the sub-tests we design and use in this study valid?**

We can begin to answer by saying that our level of error is higher than we anticipated, and that therefore they lack the level of response validity we wanted to achieve.

This leads us to consider another of our research questions:

- What modifications can we introduce to improve the sub-tests?

Without doubt, if we look at the data derived from employing the SPEARMAN-BROWN PROPHECY FORMULA (Table 18—9) we can see that all of our sub-tests could be lengthened to achieve greater reliability and thus reduce the S_e. We report on this in more detail below.

18.1.2 ***PREDICTIVE VALIDITY (CONDITION 6)***

In the following sections, we look at the correlation coefficients derived from comparing the pairs of instruments, and we comment on the strengths of the relationships in both Cohort A and Cohort B.

The correlations in question are based on the use of the PEARSON PRODUCT-MOMENT FORMULA, and KENDALL'S TAU$_B$ FORMULA. In each case, we indicate the statistical significance of the correlation with reference to the tables of statistical significance reproduced in Silver (1997:102), and Siegel and Castellan (1988), and to a 5% level of significance. Shading in the tables presented in this chapter indicates the correlations that are statistically significant, and we will discuss each of these individually.

We will later comment on the correlations between other instruments in the study, and draw further conclusions from the data we have generated.

18.1.2.1 *Cohort A*

In this section we will discuss the six correlations arising from the comparison of these data, three of which are statistically significant, although they all fail to reach our target standard (Table 18—3).

DISCUSSION

Table 18—3 Cohort A Predictive validity correlation coefficients

	LA	*LB*	*A-B*	*B-A*
LA	1.00			
LB	0.08	1.00		
A-B	0.13	0.45	1.00	
B-A	(0.09)	0.25	0.32	1.00

Language A reading skills and Language B reading skills

As would be predicted, there is virtually no correlation between these skills in the two languages. This is perhaps indicative of the low level of LB proficiency candidates have when compared to that in their mother tongue: few, if any, approach LB native speaker standards. At a higher level of proficiency, we would expect a greater level of skill transfer from LA to LB. This would reflect the fact that reading skills are taught explicitly in the FL classroom, but not in the mother tongue, where it is assumed — perhaps erroneously — that all individuals have acquired these.

Language A reading skills and General translation A-B

There is a low positive correlation here, which does not reach the level of statistical significance. We would suggest that this is because the criteria that predominate in the correction of the General translation A-B examination are heavily oriented towards the target language. Most candidates will have a more than sufficient command of their mother tongue to enable them to decipher the content of the source text, and the elements that discriminate between candidates are those found in their ability to express themselves in the target language.

Language A reading skills and General translation B-A

The fact that these tests should produce a negative correlation of - **0.09** perhaps indicates the differences in the nature of the skills being evaluated. The active skill of written production is that which dominates, to the extent that LA reading ability is an irrelevance.

Language B reading skills and General translation A-B

Here we have a large sample (n>50) and we have attained a positive correlation of **0.45**. This fails to reach the minimal requirement of our research design, but is statistically significant and, if it were corroborated by the results of Cohort B (below), would counter the poor results described by the negative correlations we have discussed above.

Language B reading skills and General translation B-A

The logic of our study implies that the relationship between LB reading skills and General translation from LB into LA should be stronger than the complementary relationship we have just described. Unfortunately, this is not the case. The correlation of **0.25** — while it is statistically significant — is weaker.

The explanation for this may well be found in unknown external variables, which have intervened in the development of the candidates over the two years of the study. Alternatively, it may lie in elements of the entry test instruments themselves, possibly in the content validity of the test items with respect to the target performance in translation. We believe we have been testing reading skills, but we may well been testing other skills in addition, which in some way have marred our results. On the other hand, it may even lie in aspects of the target performance testing in General translation B-A.

General translation A-B and General translation B-A

Finally, we come to the relationship between the two examinations in translation, which complement each other. The low positive correlation we find here, **0.32** indicates the common and disparate elements of the two subjects. The common elements lie in the practical aspects of translation, taught by two teachers who have worked in close harmony over a number of years. The divergences lie in terms of the directionality of the translations, or the criteria the teachers apply, or the relative weighting they give to each criterion.

CONCLUSIONS

Two years into the Translation and Interpreting studies program, there are clearly a wide range of criteria, both linguistic and methodological, which appear to have more bearing on the correction of translation examinations than reading skills do. Has the anecdotal emphasis on the importance of LB proficiency at the entry point — as a gatekeeper — given way to an emphasis on written production in LA and in LB, two years later? The apparent contradiction in results may have part of its origin in variation in criteria.

18.1.2.2 *Cohort B*

In this section, we will look at a set of data similar to that we have just discussed. However, there are certain important differences. The first lies in the absence of data for the Language A sub-test taken by Cohort B, which we referred to in Chapter

17. As we pointed out there, the circumstances that gave rise to this lack of data may well have influenced other elements of the sub-tests and may have affected the quality of the data we will now discuss.

A second difference between these two cohorts lies in the modifications to the test specification that were made. The most important of these involved the introduction of a set of Error identification and Error correction items in an attempt to test aspects of written production through an objectively marked task. The influence of this task should be most evident in the correlation between the LB sub-test and the General translation A-B examination, that is, translation into English.

Not one of the correlations that we have calculated from these data is statistically significant, and one is actually negative. However, this and all of the non-significant, low positive correlations needs careful analysis, particularly as these correlations are lower than those produced by Cohort A. Specifically, the negative correlation between the Language B sub-test and General translation B-A, where we would expect the highest positive correlation, requires serious thought. All of this discussion needs to be carried out in the light of the fact that the LB sub-test concerned, in03, produced the highest split-half reliability coefficient (0.619). This suggests that flaws in the test as a test instrument are less likely to be the cause, and that other factors must have intervened.

Table 18—4 Cohort B Predictive validity correlation coefficients

	LB	*A-B*	*B-A*
LB	1.00		
A-B	0.14	1.00	
B-A	(0.07)	0.21	1.00

DISCUSSION

Language B reading skills and General translation A-B

The correlation coefficient of **0.14** is very poor when compared with that attained by candidates in Cohort A (**0.45**). It seems to indicate major changes in the levels of ability developed in the candidates over the two year period, and would require a detailed study of individuals and of their academic and personal development over that period in order to establish any firm explanation.

It was our intention, by including the Error identification and Error correction items to measure written production in the LB, and we assumed that this would have an influence on the predictive validity by raising the level of the correlation with General translation into the LB. However, it does not appear to have done so. Perhaps, again, this is an indication of the elements of criteria used in the correction of the General Translation A-B examination, where accuracy and precision give way to aspects of style and maturity in written production.

At this point, we believe it is important to report that informal feedback from all of the teachers of translation who taught both Cohort A and Cohort B during their studies in the Faculty, coincided in that they believed the latter to be much weaker in LB English. This is reflected in the average grades attained by these students up to their final year. Only at the end of their studies did they begin to obtain grades comparable to their predecessors. These views take us beyond the scope of our research, but point to further lines of research of interest.

Language B reading skills and General translation B-A

As we mentioned in the introduction to this section, the negative coefficient produced by this correlation has confounded our hypothesis of the link between these two measures. A coefficient of **–0.07** can only lead us to speculate as to the nature of the instruction and the learning undergone by the candidates. During the two years between the administration of these test instruments, the range of other variables that may have intervened to a greater or lesser extent is unknown.

We do not discount the fact that the test instrument, sub-test in03, may have been deficient in some way or another. However, we feel that as it produced the highest reliability coefficient of all of the papers, and as the item analysis (Chapter 17) was reasonably good, it is less likely to have been the cause of this result. The S_e was 1.99, but the unweighted score was based on 30 points, making it relatively good.

To conclude, the only explanation we can find is that other variables, such as individual language learning processes may have influenced candidates' progress to the extent that the correlation should prove negative. Perhaps candidates who had gained entry to the degree program and found that their level of LB was relatively poor enrolled in complementary language classes in order to improve. However, we find it hard to believe that this particular variable could be the only explanation.

General translation A-B and General translation B-A

A comparison of the circumstances and instruments involved in teaching General translation A-B and B-A to Cohort A and Cohort B indicates few differences. The teachers involved were the same ones and the core content and materials used during the

General translation classes were unchanged. However, in the case of Cohort A, the General translation courses were taught in consecutive semesters. General translation B-A was taught from October to February, and A-B was taught from February to June. Each subject was examined at the end of the corresponding semester. The following year, Cohort B was taught B-A and A-B simultaneously from October to June, with examinations held in the June session. This variable may have influenced the statistically non-significant coefficient of Cohort B (**0.21**).

As in the case of Cohort A (**0.32**), we have here a low positive correlation. We expect that the causes of this lie in the same area as in the case of Cohort A. That is, that these would derive from the inherent differences of the directionality of the subjects, and the differences of criteria and weighting of criteria applied by the two teachers involved. Alternatively, this lower coefficient may be a consequence of the changes in the duration of the teaching period that we have described earlier. Alternatively, they may be the consequence of a combination of these factors.

18.1.2.3 *Conclusions*

Predictive validity is the criterion by which we have chosen to evaluate our hypothesis that

Reading skills provide a measure of aptitude for translation

We believe that in view of the results generated from our longitudinal study of Cohort A, we are in a position to reject the null hypothesis:

No direct relationship exists between reading skills and aptitude for translation

We do so in the knowledge that the incomplete data generated by Cohort B contradict that provided by Cohort A. However, we have sufficient confidence in the results of sub-test in03, to believe that these are a consequence of other variables outside of our control.

However, we are not able to support our initial belief that the language in which the reading skills are tested would be evident in the directionality of the translation used to test predictive validity. In other words, we have to accept the null hypotheses that

No direct relationship exists between LA reading skills and translation from LA into LB
No direct relationship exists between LB reading skills and translation from LB into LA

We now move on to the cross-sectional part of our study, and look at the degree to which our test instruments achieved validity against a further series of performance criteria based on the CTT classification of types of validity. In Table 18—5 we present the summary of criteria, means of testing and parameters. In the following sections, we will discuss these data with reference to Cohorts A and B.

18.1.3 *CONCURRENT VALIDITY*

Table 18—5 How did we intend to test concurrent validity? What was our target?

Means of testing	Correlation coefficients between LA and LB sub-tests and OPT, University entrance test, and Secondary school average
Target parameters	≥0.65 at 5% significance

The concurrent validity of any test, as defined in Chapter 7, is measured in terms of the correlations between the scores it generates and those attained in other tests of the same knowledge or skill(s). In the cross-sectional part of our study, we have not administered any parallel tests of LA or LB reading skills. However, we have administered an English language placement test, the Oxford Placement Test (OPT) (Allan 1991), and we have introduced data on the candidates' University entrance examination (Univ) and Secondary school average scores (Ssch).

In Chapter 3, we described the results of administering the OPT to Diploma course students in an attempt to establish their LB proficiency level. Later, in Chapter 8, we detailed the components of the test, and indicated the overlaps between its content areas and our more narrowly defined reading skills sub-test. At that point, we stated that we had decided to use it as one means of validating our work.

Furthermore, in Chapters 3 and 4, we have encountered the question of the relationship between general intelligence and aptitude. Both Bossé-Andrieu (1981) and J.B. Carroll (1981) make the point that there are connections between the two, and the former includes measures of secondary school performance in her statistical analysis. Similarly, the bulk of research into aptitude documented by Spolsky (1995) and discussed in Chapter 4, points to the 0.20-0.60 range as being that most commonly found. Therefore, we conclude that this aspect of candidate performance could have a bearing on our research. The University entrance examination entails tests in LA Spanish and LB English which clearly overlap with our sub-tests in a similar, although less precise manner, as the OPT does.

Our prediction, therefore, is that there should be some low, positive correlations between these measures and those of the LA and LB sub-tests. Table 18—6 reveals the strengths and natures of these correlations, which we discuss below. If these correlations prove strong then we will have to consider the practicality of using one or other of these tests in the place of our own. We will deal with this in a later section when we discuss the issue of test practicality.

18.1.3.1 *Cohort A*

Table 18—6 Cohort A Concurrent validity coefficients

	Univ	*Ssch*	*LA*	*LB*	*OPT*
Univ	1.00				
Ssch	0.75	1.00			
LA	0.21	0.09	1.00		
LB	0.27	0.11	0.08	1.00	
OPT	0.28	0.08	0.27	0.35	1.00

UNIVERSITY ENTRANCE EXAMINATION AND SECONDARY SCHOOL AVERAGE

As was to be expected, the correlation between these two scores was **0.75** which clearly shows a high level of similarity between what are, after all, measures of essentially the same elements, but with criteria applied by two different although overlapping collectives.

UNIVERSITY ENTRANCE EXAMINATION AND LANGUAGE A READING SKILLS

The teachers of Language A within the Faculty of Translation and Interpreting at the University of Granada argue that the administration of a separate test of LA reading skills is unnecessary and in recent years, such a test has been dropped from the entry procedures. Their position is based on the perceived similarities between the reading skills sub-tests in our model and those found in the University entrance examination. A correlation of **0.21** between these two measures is statistically significant, and their position is supported by these figures.

UNIVERSITY ENTRANCE EXAMINATION AND LANGUAGE B READING SKILLS

The correlation we find here (**0.27**) is, surprisingly, slightly higher than that for Language A. As a low positive correlation it does indicate some sort of connection between performance on the two measures and might be offered as an argument for the abolition of a Language B entry test, too.

One general comment, which may be relevant to the interpretation of the three relationships, is the fact that candidates took all three examinations in a short period of time at the end of months of intensive preparation. It is therefore likely that their level of examination preparedness may also have influenced the scores.

UNIVERSITY ENTRANCE EXAMINATION AND THE OPT

We did not expect this relationship to be significant. As a general measure of English language proficiency, the OPT gives a low positive measure of correlation (**0.28**) with *selectividad* which is logical in as much as there is an English language component to the examination. However, the comparatively small sample renders the results statistically insignificant.

SECONDARY SCHOOL AVERAGE AND ALL OTHER MEASURES

None of the correlations between the Secondary school average and the other instruments in our study are statistically significant. It may well be considered obvious that this was likely to occur. However, the purpose of including these correlations was to follow the line of investigation that appears in Bossé-Andrieu (1981; see Chapter 3) and that is clearly present in all of Carroll's writings, namely that some sort of relationship exists between aptitude and the *g* factor, i.e. general intelligence. Bossé-Andrieu links these in her research and finds a correlation of 0.254 between Secondary school average and Translation B-A, although she does not state the sample size, nor does she affirm the statistical significance of the relationship. If we are to give her research the statistical benefit of the doubt, then our results firmly contradict this relationship. From this, **we conclude that a measure of aptitude other than the secondary school average *is* required.** If we combine these data with those derived from the University entrance examination, we must conclude that an aptitude test of one sort or another does serve a purpose in the selection process.

LANGUAGE B READING SKILLS AND THE OPT

In this instance we have a low positive correlation of **0.35** which, if we consider the relatively few individuals in the OPT sample (n >30) is remarkably good. The strength of the relationship would need to be corroborated by other studies in order for us to make substantial claims for our test instrument, but it is nonetheless a satisfying result. We can state, with a minimal degree of certainty, that the test has measured skills that probably overlap with those measured by the OPT.

18.1.3.2 *Cohort B*

Table 18—7 Cohort B Concurrent validity coefficients

	Univ	*Ssch*	*LB*	*OPT*
Univ	1.00			
Ssch	0.46	1.00		
LB	0.12	0.08	1.00	
OPT	0.05	0.03	0.45	1.00

UNIVERSITY ENTRANCE EXAMINATION AND SECONDARY SCHOOL AVERAGE

The correlation between these scores (**0.46**) is surprisingly low when compared to that for Cohort A (**0.75**). Perhaps, if we contrast figures for three samples we can put these figures into perspective. By "All applicants", we mean all of those who took the entry test. "Candidates accepted", describes those candidates who were offered places on the degree program in Translation studies; and "Candidates rejected" were those who were not offered places as a result of the entry test and the combination of their score on this test with their University entrance test score.

Table 18—8

	Correlation coefficient	*Univ average*	*Ssch average*
All applicants	0.72	6.88	7.25
Candidates accepted	0.46	7.92	8.12
Candidates rejected	0.56	6.51	6.94

Regardless of the level of the correlation, the level of ability demonstrated by the University entrance examination (Univ) and Secondary school averages (Ssch), respectively, indicate a sample of high general ability. In Bossé-Andrieu's terms, these are candidates whose general level of academic ability should enable them, given time, to acquire the language skills necessary regardless of their initial language level. The lower coefficient recorded for this sample may merely be due to the fact that the sample was much smaller (n>60) than that of Cohort A (n >150).

One point of comparison available to us here are data on the diploma intake students from 1992-93 (Table 18—9). These show a slightly higher level of correlation between the two measures, but rather lower average scores.

Table 18—9

	Correlation coefficient	*Univ average*	*Ssch average*
Candidates accepted	0.79	7.21	7.65

UNIVERSITY ENTRANCE EXAMINATION AND LANGUAGE B READING SKILLS

The very low, positive correlation (**0.12**) here is such that it is difficult to comment with any degree of certainty. The correlation is weaker than that for Cohort A (**0.27**) and, as in the case of the previous correlation, this may be due to the lower sample size, or the circumstances surrounding the administration of the test.

UNIVERSITY ENTRANCE EXAMINATION AND THE OPT

The pattern of lower correlations for Cohort B (**0.12**) when compared to Cohort A (**0.28**) is continued in these data.

SECONDARY SCHOOL AVERAGE AND ALL OTHER MEASURES

Similarly, the coefficients recorded between this set of measures and the Secondary school average show an absence of any relationship, and are also slightly lower than those for Cohort A.

LANGUAGE B READING SKILLS AND THE OPT

This particular coefficient runs counter to all of the others in the set of values recorded for Cohort B in that it is stronger than that recorded for Cohort A, and is statistically significant. The value of this figure is that it confirms the information offered by the results provided by Cohort A, and indicates that there is a clear overlap of knowledge areas and skills between the two instruments. Furthermore, it would appear that the overlap is greater in the case of sub-test in03, in which we introduced the Error identification and Error correction items, than it was for sub-test in02.

The conclusion that we are led to by these strengthens our belief that the criteria used to evaluate General translation A-B lay more emphasis on aspects of written production in English other than those of accuracy and precision. Alternatively, it may be that these aspects have been adequately taught to candidates during the two years of study before the General translation exam.

18.1.3.3 *Conclusions*

The concurrent validity of our tests would seem to have been established by the strength and nature of the correlations with the OPT test. The relative weakness of the correlations against the wider ranging measures indicate that our sub-tests do, in fact, test discrete sets of skills. It is, therefore, to our analysis of construct validity that we next turn in order to assess whether these discrete sets of skills are those we believe them to be.

These results are in line with those found by Clapham (1975, reported in Alderson 1984:13.

> *The study concluded that the best predictor of reading ability in a foreign language was not reading ability in the mother tongue, but rather proficiency in the foreign language.*

The correlation coefficient in her study ($r = 0.67$) was much higher than that which we generated, but Alderson points out that the results were influenced by "text effect", in that LB proficiency was of greater importance when candidates were working with an "easy" text. Reading ability was of greater importance when they were working with a more "difficult" one.

18.1.4 *CONSTRUCT VALIDITY*

Table 18—5 How did we intend to test construct validity? What was our target?

Means of testing	Correlation coefficients between LB sub-tests and OPT
Target parameters	≥0.65 at 5% significance

The theoretical construct that we are attempting to measure is that of reading skills. Objective tasks, such as the multiple-choice questions and most of the other tasks we have used, are largely based on reading skills. Similarly, the grammar and listening components of the OPT are wholly based on the reading of prompts. Therefore, we consider that the construct validity of the sub-tests can be measured through the strength of the correlations between the LB and the OPT scores, shown in Table 18—8.

As these correlations are similar in strength for the two cohorts, and both are statistically significant, we consider that a good degree of construct validity has been proven. We have failed to achieve our target parameters, but the results are in line with those of researchers in FL aptitude. As we have discussed above, part of this overlap is almost certainly in the field of reading skills; a further part, we suggest lies in the area of written accuracy.

Table 18—8 Language B reading skills and the OPT. These correlation coefficients indicate concurrent, construct, and criterion validity.

Cohort A	0.35
Cohort B	0.45

18.1.5 *CRITERION VALIDITY*

Table 18—9 How did we intend to test criterion validity? What was our target?

Means of testing	Correlation coefficients between LB sub-tests and OPT
Target parameters	≥0.65 at 5% significance

In our study, criterion validity is measured in terms of the strength of the correlation between the LB sub-tests and the OPT, this being the externally validated measure of English proficiency that we have adopted for the purpose. The OPT tests administered to both cohorts produced statistically significant coefficients with the LB sub-tests (Table 18—8), which demonstrate an effective overlap between the test instruments even though they are below our target value. This would also seem to prove the criterion validity of our instruments, meaning that as tests of some areas of English proficiency they are valid.

Criterion validity is important to our study as it enables us to make appropriate comparisons between these cohorts and the students who made up the earlier Diploma course population. As we have said above, this provides a consistently strong correlation in the two sets of data (Cohort A r = **0.35** and Cohort B r = **0.45**).

18.1.5.1 *Conclusions*

We can conclude, therefore, that the hypothesis

Our LB reading skills sub-tests measure skills and areas of knowledge that form part of general English language proficiency

can be accepted. And we reject the null hypothesis

Our LB reading skills sub-tests measure skills and areas of knowledge other than those that form part of general English language proficiency

In addition, the data on candidate performance in the OPT leads us to confirm another of our initial hypotheses, namely that

The introduction of an entry test that includes an LB sub-test will raise the standard of general LB proficiency
The introduction of an entry test that includes an LB sub-test will have no effect on the standard of general LB proficiency

In this case, we offer a set of comparative results (Table 18—10), in percentages, which indicate a general improvement in LB proficiency over three years: the final Diploma intake 1992-93, and the two degree course cohorts A (1993-94) and B (1994-95).

The general trend over the three years has seen an upward shift in terms of the percentages of candidates in each band, most striking of all in Cohort A. The mean score rose by 17 out of 200, which just failed to take the mean grade into a higher band. Cohort B saw a lapse in performance as far as the mean grade was concerned, and a wider spread of grades overall. The two candidates who achieved Pre-elementary scores did so because they failed to complete all parts of the paper, but we have been unable to determine whether this was due to lack of time or oversight.

Table 18—10 A comparison of Language B proficiency based on the OPT scores for three generations of students. The descriptors are taken from Allan (1990).

Descriptor	*Diploma (92-93)*	*Cohort A (93-94)*	*Cohort B (94-95)*
Mean score (out of 200)	*140*	*157*	*144*
N =	*115*	*53*	*50*
Advanced to near-native	0%	2%	1.9%
Upper intermediate to advanced	10.4%	46%	34%
Intermediate to upper intermediate	44.3%	40%	30.2%
Lower intermediate to intermediate	34.8%	12%	22.6%
Post-elementary to lower intermediate	8.7%	0%	7.5%
Elementary to post-elementary	1.7%	0%	0%
Pre-elementary	0%	0%	3.8%

18.1.6 *CONTENT VALIDITY*

Table 18—11 How did we intend to test content validity? What was our target?

Means of testing	Expert opinions offered by members of the examination board
Target parameters	Reference to the test specification

This is an element of our tests that is not measured objectively but, rather, by the subjective evaluation of the members of the Examination board who reviewed and revised the sub-test papers with reference to the written specification. In the previous chapters, where we have presented the results of each of the seven sub-tests, we have commented on their general approval of the papers, and on modifications we have made to the sub-tests based on their suggestions.

We do not mean to accept the content validity of the tests as proven, but in the light of our research, we feel that the null hypothesis cannot be rejected. It is clear that the specialists consulted felt that test items were adequate measures of the target skills as defined in the test specification. The efficiency with which specialists are able to carry out this type of content validation is question by Alderson (1990) but affirmed by Bachman et al (1988). It is a further line of research to which we could contribute in a future study.

18.1.6.1 *Conclusions*

Our hypothesis about the content validity of our tests is that

H_1 The sample of reading skills included in each sub-test is representative of those in our initial test specification

Our tests have been "approved" by those specialists called upon to do so, and accordingly we reject the null hypothesis.

$H_{1\text{-}0}$ The sample of reading skills included in each sub-test is not representative of those in our initial test specification

However, we are aware that there are many flaws in this procedure of content validation, and believe that further research would lead to a greatly improved test specification, and an improved content validation process.

18.1.7 *FACE VALIDITY*

Table 18—12 How might we have tested face validity? What could our target have been?

Means of testing	Likert scale questionnaire
Target parameters	High levels of satisfaction

In our study, for obvious reasons, we did not consider it appropriate to discuss with candidates the experimental nature of the entry tests that they had taken. Any information on the face validity of the papers was gathered on an ad hoc basis in informal conversation in the years following the entry tests. This aspect of validity has been deliberately ignored from the formal research perspective, and we are well aware that it is a failing in our research design. Face validity has enormous consequences in the overall context of test administration and we consider it important to follow up this line of investigation in the near future.

Had we wished to conduct a survey, we would have followed a model similar to that used by Weir (1983). This would have involved administering a Likert scale questionnaire immediately after the test session. The questionnaire could have been followed up at a later point by one-to-one interviews with selected individuals to expand on their questionnaire responses.

18.2 Are the sub-tests we design and use in this study reliable?

In this section, we compare the levels of reliability achieved by the four tests that form a part of the longitudinal part of our study. These tests are included in the sets of measures that we have chosen to use in order to validate the test instruments. They are sub-tests es01.2, es02, and in02, all of which were administered to the candidates in Cohort A, and sub-test in03, which was administered to those in Cohort B.

From a technical point of view it is evident that all of the sub-tests used in the full sessions in our study failed to achieve our target of >0.9 for reliability. As Table 18—9 shows, the highest coefficient was achieved by the fourth of these, which represented a change in the test design. This involved the introduction of a set of error identification and error correction items to complement the multiple-choice questions and the removal from the specification of the matching exercise and the True/False/Don't know items. We also modified the MCQs by reducing the number of options from five to four.

Table 18—9 Summary of reliability coefficients

Sub-test	*Cohort*	*Split-half average*	*Extra items*	*Prophecy*
es01.2	A	0.592	+6	0.947
es02	A	0.451	+10	0.903
in02	A	0.395	+14	0.948
in03	B	0.619	+9	0.991

These technical modifications in the test specification, combined with the experience we had gained in test writing and administration, perhaps brought about the slightly higher reliability coefficient. It is important to note that sub-test es01.2 — a revised version of a trial test — attained the second highest of the reliability coefficients. This lends support to the widely held view that tests should be trialled, and revised in the light of trial results, as extensively as possible.

However, we must conclude that the individual sub-tests were not sufficiently reliable. Candidates, who hypothetically, might take the same sub-tests again on a subsequent occasion, would perhaps achieve different scores due to aspects of the test instrument, rather than due to any element that might reflect on their knowledge or skills.

The only straw we can clutch at to maintain a degree of faith in the value of results derived from these tests is to re-state the fact that we have been measuring the sub-tests as isolated instruments. From the candidates' perspective, each sub-test is one only in a battery of four, including the listening sub-tests for interpreting described in the test specification in Chapter 8, but which are beyond the scope of our research.

18.3 Does the data gen erated in our research lead to any other significant conclusions?

In the next section, we will look back at the full set of correlations derived from the set of data we have accumulated and attempt to interpret in depth a number of correlations which have to do with the practicality of our test instruments.

18.3.1 *CAN READING SKILLS BE USED AS AN EFFICIENT MEASURE IN OUR REAL-WORLD SITUATION?*

The practicality of our tests is one of the essential issues that has to be established by our analysis of these results. If it can be proven that satisfactory correlation coefficients are produced by the use of tests other than the specific entry test then this has no academic purpose whatsoever. Any decision to maintain an entry test would, in that case, be a political or even economic decision that could not be justified in other terms. An entry test, under these circumstances, would be a gatekeeper levying a toll on all candidates for entrance.

In view of the data we have gathered (Tables 18—10 and 18—11), it is evident that the academic justification for an entry test is upheld. The correlations between two of the external measures that — individually or in combination — might be used to substitute for an entry test are weak enough for them to be discounted as adequate measures of aptitude for translation. None of the correlations between the University entrance examination or the Secondary school average and General translation A-B or B-A are significant, some are negative, and all of them are weak.

Table 18—10 Cohort A The practicality of using University entrance test or Secondary school average scores instead of a specific entry test

	Univ	*Ssch*	*A_B*	*B_A*
Univ	1.00			
Ssch	0.75	1.00		
A_B	0.22	(0.05)	1.00	
B_A	0.11	(0.14)	0.32	1.00

Table 18—11 Cohort B The practicality of using University entrance test or Secondary school average scores instead of a specific entry test

	Univ	*Ssch*	*A_B*	*B_A*
Univ	1.00			
Ssch	0.46	1.00		
A_B	0.06	0.19	1.00	
B_A	0.00	(0.13)	0.21	1.00

If we look now at the data that corresponds to the relationship between the OPT and General translation A-B, or B-A we find a different picture (Tables 18—12 and 18—13).

Table 18—12 Cohort A The OPT as a gatekeeper in the making

	OPT	*A_B*	*B_A*
OPT	1.00		
A_B	0.67	1.00	
B_A	0.17	0.32	1.00

In the case of Cohort A, the correlation between the OPT and translation into English is strong. This would tend to reinforce the belief that the criteria of evaluation for the translation are much more inclined towards aspects of written production. Data for Cohort B are less strong, with the significant correlation being that for translation into Spanish, taking us back to the element of overlap between the OPT and General translation in terms of reading skills.

Table 18—13 Cohort B The OPT correlations with General translation

	OPT	*A_B*	*B_A*
OPT	1.00		
A_B	0.25	1.00	
B_A	0.29	0.21	1.00

18.3.2 *Conclusions*

In terms of practicality, it would seem proven that the data we have generated indicate that the use of the OPT could enhance the efficiency of our testing process.

The relationships between the commercially prepared OPT and General translation A-B and B-A are consistently stronger than the relationships between our tests and the translation examinations. Even with the low sample size (n>30) the coefficient for Cohort A can be considered "accurate enough for most purposes" (Cohen and Mannion 1989:168-169). This leads us to conclude that the criteria on which the General translation A-B examination is marked involve a number of elements present in the OPT, and therefore not specific to Translation studies, but absent from our test instruments.

As we have indicated elsewhere (Chapter 8) the OPT as we administered it involved a number of sub-tests of grammar and listening, all of which were entirely founded on the reading of item prompts and options. The OPT tests reading skills in a non-specific manner as these are not the focus of any items but, rather the means by which candidates reach the answers to items objectively testing other skills and elements of knowledge, such as grammar knowledge, and written precision.

A commercially trialled and tested examination obviously has advantages over an in-house test such as ours, and in this case, this has been proven more effective. One long-term question we need to ask must therefore be: Should we adopt a commercially available testing instrument such as the OPT instead of attempting to produce our own? While the answer may be conditioned by the practicalities of the matter — if we can buy copies of the OPT, so can potential candidates — the reality of the relationship between this measure and our target performance in translation cannot be ignored, and should be the focus of further research.

Finally, we should state that our hypothesis that

Translation studies involves a degree of specialization such that a specific entry test of aptitude is justified can be accepted, and we can reject the null hypothesis:

Translation studies does not involve a degree of specialization so great as to warrant the use of a specific entry test of aptitude

We had been led into this area by our reading of Bossé-Andrieu's work (1981), which has some grounding in Carroll's theoretical studies (1981), and which is supported by our colleagues in the Department of Spanish Philology,

However, in the case of the Oxford Placement Test, the results lead us to the opposite conclusion. We must accept that the hypothesis has been proven.

A specific test of aptitude for translation is not required as the relationships between the OPT scores are adequate measures

Consequently, we should reject the null hypothesis.

A specific test of aptitude for translation is required as the relationships between the OPT scores are insufficient measures

18.4 Item analysis

18.4.1 *MATCHING EXERCISES*

In the trial sub-tests, and in the tests administered to Cohort A, we used a matching exercise to begin each paper. These items were intended to make the papers "candidate-friendly", and the FV target was in the range from 0.8 to 1. We did not expect them to achieve our target for discrimination of [illegible]0.3, and only a few of them achieved a significant level of discrimination, but those that did were valuable contributions to test performance. Despite this, the item type was discarded when we revised the test specification for Cohort B in order to make way for the error identification and error correction items, because it contributed little to the overall test performance.

In the early tests, we found that most of these items served their essential purpose in that the FV scores they attained did fall within our target range. We learned from results of the first sub-tests that these items needed a clear and precise rubric, and that the format should include one, and probably not more than one, distractor option. Furthermore, our analysis of the performance of individual items taught us that the target skills on which they were based, skill 46 — *Scanning to locate specifically required information* — and the sub-skills derived from this, were more complex in their actual application than we had believed. In some cases, our analysis indicated that a more sophisticated combination of skills was required in order to respond correctly to these items.

From our analysis of these matching items, we encountered a further complication, in that the candidates who successfully complete all of the paper well within the time allowed have the opportunity to return to them and revise their responses. In this case, the skills being used are not scanning skills, but intensive reading skills. Therefore, for these candidates, the task fails to test our target skill.

In conclusion, we propose that a future test specification could include these items, based on the same target skills, with the proviso that the total number of items in the sub-test should add up to more than the 20 raw points used to date, and that the time limit should be calculated in such a way as to ensure that candidates did not normally have sufficient time to re-read and revise their responses to these items. As we have said elsewhere, more items would achieve higher reliability and reduce the level of the S_e. These items could then perform their original function, in a context in which few if any candidates would be likely to complete the paper and have time to re-read and revise their responses through intensive reading. The few scanning items that do achieve a level of discrimination would be able to show their worth.

18.4.2 *TRUE/FALSE/DON'T KNO W*

The performance of these items was more controversial than that of the matching exercises, although both types were discarded when we revised the test specification. This item type called into question the FACE VALIDITY of the test.

Candidates were not familiar with the item type, and they use of the third option — "Don't know", defined as "there is no information in the text" — was almost certainly responsible for the inconsistent performance of these items. While a number of these tasks produced DI scores of 0.30 and higher, demonstrating that they do discriminate well between candidates, many more led us to review our choice of key response on the grounds that too many candidates had chosen what we had predicted as an incorrect option.

When these items functioned well, it seemed that the combination of the type with skill 32.1 — *understanding relations between parts of text through grammatical cohesion devices of reference* — might prove particularly successful. However, we gathered insufficient data to prove this.

We would not propose to reintroduce this task type into future test specifications, but would prefer to trial other activities.

18.4.3 *MULTIPLE CHOICE QUESTI ONS*

In the two specifications of the test, the number of options used in the MCQ items was the only aspect that was altered. For the first tests we used five alternatives, and for the later ones we reduced this to four. The original decision was made in order to try to limit the random effect to a minimum, but the results indicated that we had failed to produce adequate options on too many occasions. This meant that candidates often chose between only two of the five responses, and that the others were therefore of no use. The feelings of most of the item writers were that it was difficult enough to produce good four-option items, and that there was no purpose in further complicating their work.

The results of the tests written according to the revised specification are not conclusive in this respect. The LA sub-test we have studied in Chapter 16, shows how difficult the test writing procedure can be, and the LB English sub-test has mixed results. However, in general it does seem that by limiting the number of options to four we lost nothing, and that quite possibly much was gained.

Three skills appeared repeatedly in the list of those items that had met our target performance parameters. These were skills 19, 20, and 22:

> *Deducing the meaning and use of unfamiliar lexical items, through*
>
> *Understanding explicitly stated information*
>
> *Understanding information in the text, not explicitly stated*

The data demonstrated that when we combined these skills with the MCQ task, we generally achieved good quality results.

We would include four-option MCQ items in any future test, and would increase the number of items, and the text length.

18.4.4 *ERROR IDENTIFICATION A ND ERROR CORRECTION*

This task type was introduced into the revised specification, which was used with Cohort B, and with candidates at the

University Alfonso X, "El Sabio", in Madrid. The results were mixed, in that many candidates achieved very low scores, and a superficial review of the sub-tests involved might well conclude that they task type was too difficult. However, if we look in more detail at the items, we find that many of them achieved very good and excellent levels of discrimination, despite their difficulty, e.g. DI 0.53.

A further surprise that comes from analyzing the performance of these items is that the Error correction items that focused on spelling and punctuation were among the best discriminators, even at a high level of language ability.

In a future test, we would include a set of items of this type as a means of testing some aspects of written production. We would also ensure that the two item types, that is Error identification and Error correction, were evaluated in their own right, and that the scores of sub-test components were not weighted against them.

18.4.5 *TARGET LANGUAGE SKILLS*

The basic principle of using discrete reading skills as the focus of our item-writing process has been vindicated by the performance of many of the items in these sub-tests, although it is clear that we have much to learn about this. In addition to the skills we have mentioned in the preceding sections, we have found that skill 34.2:

> *Interpreting text by going outside it, "reading between the lines"*

also produced good results in True/False/Don't know and MCQ items.

However, the fact that a discrete, analytical approach to item writing has been demonstrated to work, does not mean that we have come to accept this as a model of the manner in which candidates actually read. In fact, we have encountered a much more complex reading process in our subjective, introspective analysis of candidate responses. We are in no position to generalize about "the reading process", but we feel that we are able to say that many reading tasks are probably carried out by combining a number of reading skills and sub-skills. Whether these combinations are in fixed sets, is beyond the scope of our research, but we would suggest that different candidates probably use different approaches in order to reach the same answers.

18.4.6 *ITEM WRITING*

Our experience of item writing has been illuminated by the analysis of the sub-test results. In Part III Results, on a number of occasions we have described the way in which candidates' actual responses have led us to review, and sometimes alter the key responses we had chosen for the items. On other occasions, we have discarded items when candidates' responses have clearly demonstrated that these failed.

The production of valid options for MCQ items has revealed the value of seeking alternatives that echo, or (partially) repeat phrases that appear in the text. In our sub-tests, the items that have included this design feature have been among those that have performed best. All or most of the options have attracted a relatively even spread of candidates for these items.

In particular, we have learned much about the importance of peer review by the members of an item writing team, and about the perils of uncoordinated work. When trialling can be carried out, this is essential to the production of quality tests. When trialling is impossible, the value of content validation by peers is immeasurable.

18.4.7 *FACILITY VALUES AND DISCRIMINATION INDICES*

Our comments on item performance have been based on our interpretation of the objectively generated markers of facility and discrimination. In the process of analyzing so many sub-tests, we have found patterns of item behavior, which can be interpreted in particular ways. Items with a high FV can, and often do, discriminate between candidates, and these items serve the purpose of discriminating between candidates at the lower end of the ability range. Items with a low FV and a high DI discriminate well between candidates at the higher end of the ability range. The value of being able to interpret item performance in this way is that, when trialling is possible, final versions of tests can be assembled with a balance of these items, so that the test will have "something for everyone". This also means that the users of the test results are able to make informed decisions.

Part IV

Conclusions

19 Conclusiones

19 CONCLUSIONES

19.1 ¿Se justifica el uso de una prueba específica de aptitud?
19.2 ¿Lengua A o Lengua B? ¿Activa o pasiva?
19.3 ¿Análisis o síntesis?
19.4 El valor y la naturaleza de los coeficientes de correlación
19.5 ¿Se puede medir aptitud para traducir a traves de las destrezas lectoras?
19.6 ¿Cuáles son las modificaciones que podemos introducir en las subpruebas para mejorarlas?
19.7 ¿Qué modificaciones proponemos para la batería de pruebas?
19.8 ¿Cuáles son los campos a investigar en el futuro?

LA INVESTIGACIÓN que emprendimos genera un amplio abanico de preguntas e hipótesis. En este último capítulo comentamos cada uno de ellos individualmente, sin referirnos en detalle, a los datos estadísticos que hemos presentado en los capítulos anteriores.

19.1 ¿Se justifica el uso de una prueba específica de aptitud?

Uno de los elementos comunes a toda la investigación que sobre aptitud hemos estudiado es la correlación de la puntuación obtenida en busca de la validez criterial, a través de la comparación de ésta con medidas del coeficiente intelectual o de la inteligencia general. Estos estudios equiparan a éste último con las calificaciones de enseñanza secundaria o de la Selectividad. Esta dimensión de la investigación hace que surja la pregunta. ¿es necesaria una prueba específica de aptitud? Ninguna de las calificaciones a las que nosotros hemos tenido acceso — la de Selectividad y la media global de secundaria — ha generado correlaciones suficientemente altas con nuestros instrumentos como para hacernos preferirlas y abandonar nuestras pruebas. Además, las correlaciones entre estos dos instrumentos indican que los criterios que se aplican en cada uno de ellos son distintos. En las tres promociones que hemos estudiado los niveles de correlación entre la Selectividad y la media de secundaria parecen haberse distanciado lenta pero progresivamente. Todo ello contrasta con la suposición inicial de que la Selectividad sea meramente una medida de la validez concurrente de las calificaciones de secundaria. A todos los efectos, estos datos nos hacen eliminar la posibilidad de utilizar cualquiera de ellos.

19.2 ¿Lengua A o Leng ua B? ¿Activa o pasiva?

Uno de los puntos, que destacamos en nuestro repaso de los procesos de admisión a los estudios de traducción en España y en otros países, fue el debate que entendemos existe, sobre la primacía de una u otra de las dos lenguas. A ello añadimos la cuestión de la naturaleza de las destrezas que se consideran como indicativas de la aptitud para traducir. En los países francófonos, como hemos descrito antes, se tiende a valorar más la Lengua A; en España se da prioridad a la Lengua B. En ambos casos, se ha seguido el patrón de evaluar las destrezas activas. Nosotros obviamos el debate sobre las lenguas al incluir pruebas equivalentes en ambas, basadas en la misma especificación. El propio tribunal de la prueba fue el que estableció la diferencia, al hacer que la prueba en Lengua A fuese eliminatoria. Sin embargo, nosotros optamos por unas pruebas que midieran exclusivamente las destrezas pasivas o receptivas. Administramos éstas a los candidatos que formaron la cohorte A. Más adelante introdujimos una prueba más activa, o productiva, en la versión revisada que administramos a la Cohorte B.

Los resultados de nuestra investigación sobre el debate entre la Lengua A y la Lengua B quedan sin completar, ya que no conseguimos datos acerca de las pruebas de Lengua A para la segunda cohorte debido a circunstancias fuera de nuestro control. No obstante, el resultado obtenido al calcular la correlación entre la puntuación para la Lengua A, la de Lengua B, con otros instrumentos nos indican que el debate no ha concluido. Los coeficientes de correlación que hemos calculado, han resultado más bajos de los que hubimos esperado y no demuestran la relación clara entre ambos instrumentos que nosotros esperamos. Pensábamos que las destrezas de lectura en Lengua A generarían una correlación alta con la traducción desde la Lengua A hacia la Lengua B — ya que en esta dirección la lectura es esencial — y viceversa. Sin embargo, los resultados no lo han demostrado. De ello concluimos, que otros criterios de evaluación entran en la calificación de exámenes en las asignaturas de Traducción General A-B y B-A, hasta tal punto que disminuyen la importancia de las destrezas de lectura. La prueba de Lengua B en la que incluimos una actividad que medía la producción escrita fue la que consiguió el coeficiente de fiabilidad más alto de entre las subpruebas que hemos estudiado y los ítems, que diseñamos con este fin resultaron bastante buenos. Así que sugerimos que un ejercicio de producción escrita, tanto en Lengua A como en Lengua B, debería de formar parte de las pruebas específicas. Ahora bien, señalamos que no tendría por qué ser un ejercicio de resumen, sino que debería ser algo más general, un trabajo de producción escrita que posibilitara una evaluación objetiva de la capacidad de redacción de los candidatos. Esta actividad podría introducirse en la prueba además de las de identificación y corrección de errores que utilizamos en la subpruebas in03 y es04. El desarrollo y el uso de una subprueba de esta índole parece difícil de llevar a cabo, pero a nuestro parecer ayudará a iluminar el concepto de aptitud para la traducción.

19.3 ¿Análisis o síntesi s?

Nuestra investigación inicial nos condujo hacia un tercer dilema: la elección del diseño de nuestro instrumento de investigación. Encontramos que muchos centros utilizaban, y siguen utilizando, una prueba de valoración sintética basada en la redacción de un resumen en LB de un texto oral en LA. Nuestra experiencia de esta actividad en Granada fue negativa, debido a la falta de objetividad con la que los miembros del tribunal de exámenes llevó a cabo su cometido en cuanto a la corrección del examen. Dado el número de candidatos posibles que manejábamos y la importancia de la prueba específica de acceso, nosotros optamos desde el principio por un planteamiento analítico en el diseño de las subpruebas para que todo tuviera la máxima validez aparente y la máxima validez empírica posible. El modelo de diseño de la investigación que empleamos no incluyó ninguna medida de la validez aparente de las pruebas, como ya explicamos en la Parte II. No obstante, esta omisión consciente no nos exime de nuestra responsabilidad hacia la prueba y la manera en la que los candidatos lo percibieron. Además, nuestra decisión de adoptar una prueba de destrezas lectoras no significa que nos opondríamos al uso de una prueba apropiada de producción escrita. El trabajo que actualmente se lleva a cabo en la Universidad Autónoma de Barcelona (Fox 1997) demuestra la preocupación generalizada por establecer unos criterios objetivos para la evaluación de una prueba de esta índole y que debería hacer avanzar el proceso de encontrar un instrumento adecuado.

19.4 El valor y la naturaleza de los coeficientes de correlación

Establecimos el coeficiente $r \geq 0.65$ como la meta para todas las correlaciones, aunque somos conscientes de que es mayor de lo que, típicamente, se encuentra en estudios similares a lo largo de la historia de las pruebas de aptitud. Spolsky (1995) informa de muchos estudios y en la mayoría las correlaciones que se consiguen tienen valores entre 0.20 y 0.60. Casi todos nuestros resultados alcanzan niveles por debajo de nuestra meta y aparecen entre estos parámetros. Algunos son aún más bajos. Ehrman (1992), como señalamos antes, eligió $r \geq 0.20$ como meta en su estudio, ya que poseía muchos datos acerca de la homogeneidad de la cohorte que examinaba. Arriba, mencionamos los datos que tenemos sobre las notas de Selectividad y de la media global de COU, que indican un alto grado de homogeneidad entre los individuos de nuestras cohortes. No obstante, mientras nos tiente la posibilidad de creer que nuestras cohortes fueron tan homogéneas como la de Ehrman, decidimos ser más cautelosos y no seguir su planteamiento en la interpretación de nuestros resultados.

19.5 ¿Se puede medir la aptitud para traducir a través de las destrezas lectoras?

El propósito global de nuestra investigación es establecer la conveniencia de emplear una prueba de destrezas lectoras, diseñada para corregirse de manera objetiva, como instrumento para medir la aptitud para traducir.

Las limitaciones más importantes sobre el diseño de nuestra prueba fueron las que aparecen en la legislación que autorizó la prueba de acceso, donde se estipula que ningún elemento de dicha prueba puede examinar destrezas o conocimientos que formarían parte de la instrucción específica que los candidatos admitidos recibirían en sus estudios posteriores. Interpretamos este texto como una prohibición hacia el uso de cualquier actividad de traducción como tal y nos dirigimos hacia actividades similares a las que creímos que los candidatos conocerían por haberlas practicado en sus estudios en la enseñanza secundaria. Éstas no serían actividades de traducción, pero se basarían en destrezas fundamentales del proceso traductor.

Para llevar a cabo nuestro estudio, adoptamos el diseño de investigación correlacional descrito por J.B. Carroll (1981), y establecimos una serie de medidas y parámetros, como meta para evaluar la precisión de nuestros resultados. Derivamos éstos de una serie de fuentes dentro del campo de la teoría clásica de examen (Alderson et al 1995; Bachman 1990; Bachman and Palmer 1996; Weir 1983, 1990,1993; and Heaton 1988, 1999, entre otros). Diseñamos nuestro instrumento como fruto del estudio de la investigación en el campo de la aptitud para el aprendizaje de lenguas extranjeras (Carroll, J.B. 1954, 1959, 1974, 1978, 1981, 1993; Ellis 1986), y en los campos de la teoría de la traducción (Bell 1991; Reiss 1992; Lörscher 1986, 1991, 1992) y de la pedagogía de la traducción (Gile 1995, Beeby Lonsdale 1996). Estos últimos nos condujeron a la elección de las destrezas lectoras, definidas por Munby (1978), como elementos esenciales en que basar la especificación de nuestras subpruebas. Derivamos el proceso de diseño de los instrumentos de Carroll, B.J. (1980), Carroll, B.J. and Hall (1985), and Alderson et al (1995), entre otros.

Llegamos a la conclusión de que mientras las destrezas lectoras sí representan una de las variables que entran en el proceso traductor, es evidente que influyen otras variables. Por lo cual proponemos que, cualquier prueba específica de aptitud incluya un examen de elementos discretos, basado en las destrezas lectoras, además de otras subpruebas de otro tipo de destrezas y conocimientos.

Después del análisis de los resultados, pensamos que la evaluación de la calidad de producción escrita, de uso gramatical sobre todo en cuestiones de precisión lingüística, deberían incluirse en unas subpruebas equivalentes a las de destrezas lectoras. Seguimos sin encontrar prueba alguna del valor de actividades de resumen, como única prueba de aptitud para traducir y sugerimos que, este tipo de actividad integral incluye demasiados conocimientos y destrezas para evaluarlos de forma adecuada. Preferimos "desenredarlos" de manera que se pueda intentar evaluar cada destreza individualmente.

19.6 ¿Cuáles son las modificaciones que podemos introducir en las subpruebas para mejorarlas?

Estos resultados señalan la importancia de ensayar y experimentar con los exámenes ante de su uso y la del uso de subpruebas más largas (por ejemplo, las de LB podrían ser de dos o más textos, de unas 700 palabras cada uno, con unos 60 items; y una duración de 2 horas) para conseguir el grado de fiabilidad más alto y minimizar el error estándar de medida. Creemos que se justifica el uso de los ítems de elección múltiple con cuatro alternativas y las actividades de identificación de errores y corrección de errores.

19.7 ¿Qué modificaciones proponemos para la batería de pruebas?

Creemos que una batería de pruebas amplia conseguiría niveles más altos de correlación entre la prueba de acceso y los exámenes de Traducción general que usamos para comprobar la validez predictiva. Entendemos que existen pruebas suficientes en las que se evalúan otras destrezas y conocimientos en los exámenes de traducción y que éstos no se limitan a lo que se enseña a lo largo del curso en el que se ha estudiado traducción. El nivel de las correlaciones que generan las relaciones con la OPT indica claramente que éste es el caso de los exámenes de Traducción general A-B. La OPT mide la validez concurrente ya que se realizaron ambas pruebas en fechas relativamente próximas, por lo cual si incluyéramos pruebas de la precisión gramatical de los candidatos éstas tenderían a subir el nivel de las correlaciones.

19.8 ¿Cuáles son los campos que se pueden investigar en el futuro?

- Se podrían desarrollar otros estudios longitudinales acerca del desarrollo individual de los estudiantes a lo largo de los cuatro años que dura la licenciatura. En especial, se podría indagar en las fuentes complementarias de aprendizaje de LB que utilizan y las otras influencias de desarrollo personal y académico que reciben.

- La investigación comenzada por Bachman et al (1988) y Alderson (1990) acerca de la capacidad de los especialistas de acordar la validez de contenido de los ítems de las pruebas, es otro campo al que podríamos contribuir con nuestros próximos estudios.

- La validez aparente de las pruebas de acceso que se utilizan en la Universidad de Granada, además de la de los otros instrumentos de evaluación que se utilizan en la enseñanza de la licenciatura en Traducción e Interpretación, es un tema que pensamos investigar detalladamente en el futuro próximo.

- Las correlaciones altas en la relación entre la OPT y las notas de Traducción general nos conducen a la hipótesis de que en la evaluación de la traducción se incluyen elementos no específicos de la traducción sino, más bien, asociados con las destrezas en las se basa la especificación de la OPT.

- Se podrían desarrollar subpruebas en producción escrita para medir la madurez de los candidatos como autores en LA y en LB.

- Se podrían investigar las destrezas lectoras subyacentes en el proceso de resumir el contenido del texto para definir aquellas que se pueden evaluar de manera objetiva en pruebas de ítems discretos.

19 CONCLUSIONS

19 CONCLUSIONS

19.1 IS A SPECIFIC ENTRY TEST JUSTIFIED?
19.2 LANGUAGE A OR LANGUAGE B? ACTIVE OR PASSIVE?
19.3 ANALYTIC OR SYNTHETIC?
19.4 THE STRENGTH AND NATURE OF CORRELATION COEFFICIENTS
19.5 CAN APTITUDE FOR TRANSLATION BE MEASURED BY READING SKILLS?
19.6 WHAT MODIFICATIONS CAN WE INTRODUCE TO IMPROVE THE SUB-TESTS?
19.7 WHAT MODIFICATIONS CAN WE INTRODUCE TO IMPROVE THE BATTERY OF TESTS?
19.8 WHAT FURTHER RESEARCH IS NEEDED?

THE RESEARCH exercise we have undertaken has produced a wide range of questions and hypotheses. In this final chapter, we comment on each of these in turn, without referring excessively to the statistical data that we have presented in the previous chapters.

19.1 Is a specific entry test justified?

One of the common elements of the research studies into aptitude that we have looked at, is the correlation of scores produced by criterion validation of these through comparison with indicators of IQ, or general intelligence. These studies have often equated the latter with secondary school grades or University entrance examinations. The question arising from research in this direction is fundamental: is there a need for a specific test of aptitude? Neither of the two measures available to us — the University entrance examination score and the Secondary school average — has consistently provided strong enough correlations with our measures for us to abandon the use of an entry test. Moreover, the correlations between these two instruments have indicated that the criteria applied are rather different. The levels of correlation between the University entrance examination and the Secondary school average over the three years we have studied would seem to show a slow distancing of criteria between the two measures. This contrasts with our initial supposition that one is a concurrent validation of the other. This effectively rules out the use of one or other, or both of these instruments instead of a specific entry test.

19.2 Language A or Lan guage B? Active or passive?

The apparent debate over which of the two languages is of greater importance, was one of the issues arising in our analysis of the processes of entrance administration in Spain and around the world. We add this to the question of the nature of the skills considered indicative of aptitude for translation. In the francophone countries, as we have described earlier, the tendency was to value LA more highly; in Spain, the emphasis was on LB proficiency. In both cases, the pattern was to test active skills. We avoided the LA vs. LB issue in our tests by including equivalent papers in each language, based on the same test specification. The Exam board established the difference, which made the LA test eliminatory. But we opted to examine the so-called passive, or receptive skills in the sub-tests administered to Cohort A, and introduced a more active, or productive task in the revised tests, administered to Cohort B.

The results of our research into the LA/LB debate are incomplete, in as much as we were unable to obtain LA data for the second intake due to circumstances beyond our control. However, the results of correlating LA and LB scores with the other measures in the study indicate that the LA/LB debate has yet to end. The correlation coefficients recorded have been lower than could have been hoped for, and they have not shown the clear links we would have expected. It was our belief that reading skills in LA would correlate highly with translation from LA into LB — that being the direction in which LA reading skills play a crucial role — and vice versa. However, that has not proved to be the case. From this we conclude that other criteria enter into the evaluation of the General translation A-B and B-A examinations to such an extent that they reduce the relevance of the reading skills. The LB test that included a measure of written production was the most reliable of those we used, and the specific items functioned reasonably well. Accordingly, we would suggest that there must be a place for a written production exercise, in both LA and LB, in the entry test. We would point out, though, that this should not be a summary writing exercise, but more a general "essay writing" task, which would enable the objective evaluation of candidates' writing skills. This could be introduced in addition to the Error identification and Error correction tasks used in Sub-test in03. Developing and using a task of this type probably seems a tall order, but we believe it would shed further light on the aptitude of potential translators.

19.3 Analytic or synthe tic?

The third issue arising in our initial research was the dilemma over the choice of approach to the test instrument design. We found that many centers used, and continue to use, a synthetic test method based on the writing of a summary in LB, from an oral text in LA. Our experience of this particular method in Granada was negative due to the lack of objectivity with which members of examination boards carried out the testing and marking processes. Given the scale and importance of the entry test procedure, we decided early in the research design process that an analytical approach to the test design was essential in order to give the test the maximum degree of face validity, and of empirical validity, possible. In our final research design model, we did not attempt any systematic evaluation of face validity, as we have mentioned in Part II, above. However, this conscious omission did not relieve us of our responsibility for the test, and for the way in which candidates would perceive it. Furthermore, our decision to adopt a reading skills test does not mean that we would not consider the appropriate use of a test instrument designed to measure written production. The work currently being carried out at the Autonomous University of Barcelona (Fox 1997) demonstrates the widespread concern for the establishment of objective criteria for the evaluation of such test instruments and should go a long way towards offering a suitable instrument.

19.4 The strength and nature of correlation coefficients

In our research design, we established $r \geq 0.65$ as the target for all of our correlation coefficients, although we were fully aware that this figure was higher than the range typically found in similar research throughout the history of aptitude testing. Spolsky (1995) reports on a number of studies, and many of these achieve correlations within the range of $r = 0.20$ to 0.60. Almost all of our results fail to achieve our target, and fall within this latter range. Some are lower still. Ehrman (1992), as we have mentioned before, chose to accept $r \geq 0.20$ as a satisfactory result for her research, given that she knew much about the homogeneity of the cohort she was testing. We have already looked data on the University entrance examination scores and Secondary school averages of our cohorts, and these suggest a high degree of homogeneity among them. However, while it is tempting to believe that our cohorts were as homogeneous as those Ehrman worked with, we have decided not to follow her line, but to exercise more caution in the interpretation of our results.

19.5 Can aptitude for translation be measured by reading skills?

Our overall purpose in this research has been to establish the adequacy of an objectively marked test in reading skills as a means of measuring aptitude for translation.

The single most important constraint on our test design was that presented in the legislation authorizing the entry test. This stipulated that no element of the test could involve skills or knowledge that would form a part of the specific training which successful candidates would receive in their studies. Our reading of this precluded the use of any translation exercise as such, and directed us towards tasks of the type we believed candidates would be familiar with, from their secondary schooling, and which were not translation specific, but which were highly relevant in the translation process.

In order carry out our study, we adopted a correlational research design described by J.B. Carroll (1981), and established a series of measures and target parameters by which to evaluate the accuracy of our work. These we have derived from a range of sources within the field of Classical Test Theory (Alderson et al 1995; Bachman 1990; Bachman and Palmer 1996; Weir 1983, 1990,1993; and Heaton 1988, 1999, among others.) Our test instrument has been designed as a result of our reading of works in the area of general language learning aptitude (Carroll, J.B. 1954, 1959, 1974, 1978, 1981, 1993; Ellis 1986) and within the fields of Translation theory (Bell 1991; Reiss 1992; Lörscher

1986, 1991, 1992) and translation pedagogy (Gile 1995, Beeby Lonsdale 1996). These latter writers have led us to select reading skills, as defined by Munby (1978) as the essential elements on which to base the specification for our sub-tests. The instrument design process has been derived from Carroll, B.J. (1980), Carroll, B.J. and Hall (1985), and Alderson et al (1995).

It is our conclusion that while reading skills do represent one of the variables in the translation process, there are clearly a number of other variables involved. We would propose that any aptitude test used for entry to a program in Translation studies should include a discrete item examination of reading skills in addition to sub-tests in other areas.

From our analysis of results, we believe that the evaluation of the quality of candidates' written production, and of their use of grammar, particularly in terms of linguistic precision, should be included in sub-tests parallel to those in reading skills. We have yet to encounter satisfactory research that demonstrates the values of summary writing alone, and would suggest that this kind of integrated activity involves too many skills. We would suggest that these be "disentangled" in order to ensure that the individual skills are discretely tested.

19.6 What modification s can we introduce to improve the sub-tests?

Our results indicate the importance of trialling and pre-testing examinations, and of the use of longer tests. We would advocate sub-tests with two or more LB texts of 700 words or more in length, 60 items in total, and of 2 hours duration, in order to achieve maximum reliability, and to minimize the Standard error of measurement. We believe that the continued use of four-option multiple-choice questions, and items of the Error identification and correction types, is justified.

19.7 What modification s do we propose for the battery of tests?

We believe that a wide-ranging battery of tests would achieve higher levels of correlation between the entry tests and the predictive validity tests in General translation. We feel there is sufficient evidence that other skills and abilities are being evaluated in the translation examinations, and that these are not solely skills or abilities that have been taught during the year of the translation course. The levels of the correlations with the OPT clearly indicate this in the case of General translation A-B. The OPT is a measure of concurrent validity, taken at the time of entry to the course, therefore the inclusion of sub-tests that examined candidates' grammatical precision would tend to raise the level of the correlation.

19.8 What further research is needed?

- Further longitudinal research could be carried out on the development of students over the four-year period of their Translation studies. Specifically, this could look into the complementary language input they receive, and the other sources of personal and academic development they undergo.

- The work commenced by Bachman et al (1988) and Alderson 1990) into the ability of specialists to agree the content validation of test items is another area to which we could contribute with further research.

- The face validity of the entry tests used in the University of Granada, along with other measures used throughout the teaching of the degree in Translation and Interpreting is a topic that we would hope to research in some detail soon.

- The strong coefficient relationships arising from comparing the OPT with General translation scores leads us to believe that their are other elements involved in the evaluation of translation that are not specific to translation but, rather, closely associated with the target skills of the OPT.

- Written production sub-tests could be developed, in order to measure candidates' maturity as writers of both the LA and LB.

- The underlying reading skills involved in summarizing content could be researched in order to define those that could be objectively tested in discrete item tasks.

Bibliography

BIBLIOGRAPHY

Alarcón Navio, Esperanza. 1993. Traducción B-A (Francés). Notas Finales. [Typewritten marks sheet]. Available from the author at: Dept of Translation and Interpreting, University of Granada, Spain.

Alcaraz Varó, Enrique; Ramón y Denia, Jesús. 1980. *La Evaluación Del Inglés. Teoría y Práctica*. Madrid: SGEL S.A.

Alderson, J Charles. 1989. Reading in a Foreign Language: a Reading Problem or a Language Problem? In: Alderson, J Charles; Urquhart, AH, editors. *Reading in a Foreign Language*. Harlow, Essex: Longman Group UK Ltd.:1-27.

—. 1990(a). Testing Reading Comprehension Skills (Part One). In: *Reading in a Foreign Language* 6(1).

—. 1990(b). Testing Reading Comprehension Skills (Part Two): Test-Takers' Accounts. In: *Reading in a Foreign Language* 6(1).

—. 1991. Language Testing in the 1990's: How Far Have We Come? How Much Further Have We To Go? In: Anivan S, editor. Volume 2, Anthology series 25. *Current Developments in Language Testing*. Singapore: SEAMEO Regional Language Centre.

—. 1994. Metalinguistic Knowledge, Language Aptitude and Proficiency. 16th Annual Language Testing Research Colloquium. Mar 5-7. Washington DC.

—. 1997 Feb. [E-mail message to Bryan Robinson]. Available from the addressee at: Dept of Translation and Interpreting, University of Granada, Spain.

Alderson, J Charles; Beretta, A. 1992. *Evaluating Second Language Education*. Cambridge: CUP.

Alderson, J Charles; Clapham, Caroline; Wall, Dianne. 1995. *Language Test Construction and Evaluation*. Cambridge: CUP.

Alderson, J Charles; North, Brian, editors. 1989. *IATEFL Language Testing Symposium*. Developments in English Language Teaching. Bournemouth: IATEFL.

Alderson, J Charles; Urquhart, AH, editors. 1989. *Reading in a Foreign Language*. London: Longman Ltd.

Alderson, J Charles; Wall, Dianne. 1993. Does Washback Exist? *Applied Linguistics* 14:115-29.

Allan, Dave. 1990. *Oxford Placement Test 1.* Oxford: OUP.

—. 1992. Oxford Placement Test 2. Oxford: OUP.

ALTE Members. 1998. *Multilingual Glossary of Language Testing Terms.* Studies in Language Testing 6. Cambridge: CUP.

[Anonymous]. 1973. L'Enquête E.P.T.I. *Traduire* 76:10-19.

[Anonymous]. 1991. *Programas E.U.T.I. 1990-1992.* Granada: I.C.E., University of Granada.

[Anonymous]. 1992. Royal Decree Nº 1060/1992. In: *Boletín Oficial del Estado* :228.

[Anonymous]. 1993 Mar 6. La Factura Del Campo. *ABC.* Editorial:17.

Argüeso, Antonio. 1994 Jul. [Personal communication to Bryan Robinson.]

Aspinall, Patricia, editor. 1995. *Testing Newsletter.* Whitstable, Kent: IATEFL SIG on Testing.

—. 1998. *Testing Newsletter.* Whitstable, Kent: IATEFL SIG on Testing.

Association De Professeurs De Français Des Universités Et Collèges Du Canada (APFUCC). Date unknown. *Round Table Conference of APFUCC:* [Typewritten report] Available from Jacqueline Bossé-Andrieu, Faculty of Arts, University of Ottawa, Canada.

Bachman, Lyle F. 1990. *Fundamental Considerations in Language Testing.* Oxford: OUP.

Bachman, Lyle F; Palmer, Adrian S. 1996. *Language Testing in Practice.* Oxford: OUP.

Baker, David. 1989. *Language Testing: A Critical Survey and Practical Guide.* London: Edward Arnold.

Baker, Rosemary. 1997. *Classical Test Theory and Item Response Theory in Test Analysis.* Lancaster: CRILE/International Language Testing Association.

Barnes, Nigel. 1996. Entry Points. Workshop Report: Bias in Multiple-Choice Tests. In: *Young Learners, Testing and Research SIGs*. Whitstable, Kent: IATEFL:56-60.

Barr, Pauline; Clegg, John; Wallace, Catherine. 1981. *Advanced Reading Skills*. Harlow, Essex: Longman Group Ltd.

Baxter, Andy. 1997. *Evaluating Your Students*. London: Richmond Publishing.

Beaugrande, RA de; Dressler WU. 1981. *Introduction to Text Linguistics*. Harlow, Essex: Longman Group Ltd.

Beeby Lonsdale, Allison. 1996(a) Mar 18 [Letter to Bryan J Robinson]. Available from the addressee: Dept of Translation and Interpreting, University of Granada, Spain.

—. 1996(b). *Teaching Translation From Spanish to English: Worlds Beyond Words*. Ottawa: University of Ottawa Press.

Bendazzoli, G; Escalante G. 1992. From 'Real Life' Problems to Research. *Forum* 30(1):16-20.

Black, TR. 1993. *Evaluating Social Science Research. An Introduction*. London: Sage Publications Ltd.

Booth, Wayne C 1961. *The Rhetoric of Fiction*. Chicago: University of Chicago Press.

Bossé-Andrieu, Jacqueline. 1981. L'Admission Des Candidats Aux Écoles De Traduction. In: Delisle, Jean, editor. *University of Ottawa Quarterly: L'enseignement de l'interpétation et de la traduction*. 3(51):465-76.

—. 1996 Apr 4. [Letter to Bryan J Robinson] Available from the addressee at: Dept of Translation and Interpreting, University of Granada, Spain.

Bowker, Lynne, et al., editors. 1998. *Unity in Diversity? Current Trends in Translation Studies*. Manchester: St. Jerome Publishing.

Campagna, Louise; Dionne, Jean-Paul. 1981. Aptitudes, Intérêts, Et Réussite Scolaire En Traduction: Étude Longitudinale. In: Delisle, Jean, editor. *University of Ottawa Quarterly: L'enseignement de l'interpétation et de la traduction*. 3(51):477-93.

Canale, Michael. 1983. On Some Dimensions of Language Proficiency. In: Oller, JW, editor. *Issues in Language Testing Research*. Rowley, Mass.: Newbury House.

Carrell, P L; Devine, J; Eskey, DE, editors. 1988. *Interactive Approaches to Second Language Reading Tasks*. Cambridge: CUP.

Carroll, Brendan J. 1980. *Testing Communicative Performance*. Oxford: Pergamon Press.

Carroll BJ, Hall PJ. 1985. *Make Your Own Language Tests*. Oxford: Pergamon Institute of English.

Carroll, John B. 1954. Notes on the Measurement of Achievement in Foreign Languages. [Manuscript]. Available from Bryan J Robinson at: Dept of Translation and Interpreting, University of Granada, Spain.

—. 1974. The Aptitude-Achievement Distinction: The Case of Foreign Language Aptitude and Proficiency. In: Green, DR, editor. *The Aptitude-Achievement Distinction*. Monterey CA: CTB-Macmillan/McGraw-Hill, 286-303.

—. 1978. Linguistic Abilities in Translators and Interpreters.In: Gerver, David. Sinaiko, H Wallace, editors. *Language Interpretation and Communication*. New York: Plenum Press:119-29.

—. 1981. Twenty-Five Years of Research on Foreign Language Aptitude. In: Diller, KC, editor. *Individual Differences and Universals in Language Learning Aptitude*. Rowley, MA: Newbury House:83-118.

—. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies.* Cambridge: CUP.

Cattell, RB; Butcher, AJ. 1968. *The Prediction of Achievement and Creativity*. Indianapolis/New York: Bobbs-Merrill Co Inc.

Clapham, Caroline. 1975. *Test of English for Adult Learners (TEAL)*. University of Edinburgh.

Clapham, Caroline; Wall, Dianne, editors. 1994. *Language Testing Update*. Lancaster: CRILE/International Language Testing Association.

—. 1995. *Language Testing Update*. Lancaster: CRILE/International Language Testing Association.

—. 1996. *Language Testing Update*. Lancaster: CRILE/International Language Testing Association.

Clouse, Barbara Fine. 1997. *Working It Out: A Trouble-Shooting Guide for Writers*. 2nd ed. New York: McGraw Hill.

Cohen, AD. 1984. On Taking Tests: What the Students Report. *Language Testing*. 1(1):70-81.

—. 1980. *Testing Language Ability in the Classroom*. Rowley, Mass.: Newbury House.

Cohen, L; Holliday, M. 1982. *Statistics for Social Scientists*. London: Harper Row.

Cohen, L; Mannion, L. 1989. *Research Methods in Education*. London: Routledge.

Consejo de Universidades. 1988. *Reforma De Las Enseñanzas Universitarias. Título: Licenciado En Traducción e Interpretación*. Madrid: Ministerio de Educación y Ciencia.

Council of Biology Editors, Style Manual Committee. 1994. *Scientific Style and Format. The CBE Manual for Authors, Editors, and Publishers*. 6th ed. Cambridge: CUP.

Creswell, JW. 1994. *Research Design: Qualitative & Quantitative Approaches*. London: Sage Publications Ltd.

Crocker, Linda; Algina, James. 1981. *Introduction to Classical and Modern Test Theory*. Florida: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16:292-334.

Culler, Jonathan. 1983. *On Deconstruction*. London: Routledge.

Davies, Alan; Upshur, John, editors. 1995. *Language Testing*. 12. London: Edward Arnold.

—. 1996. *Language Testing*. 13. London: Edward Arnold.

Deschamps, Roger. 1981. [Unpublished report.] Available from Bryan Robinson at: Dept of Translation and Interpreting, University of Granada, Spain.

Dilts, Robert. 1994. *Effective Presentation Skills*. California: Meta Publications.

Dossena, Marina. 1994. Objectively-Scored Placement Tests in a Communicative Approach. *Testing Newsletter* Whitstable, Kent: IATEFL SIG on Testing:2-5.

Ehrman, Madeline. 1994. Round Table Discussion Session: Testing Aptitude Testing: The MLAT . 16th Annual Language Testing Research Colloquium. Mar 5-7. Washington DC. Available from the author at: Foreign Service Institute, US Department of State, 1501 S Randolph St, Arlington VA 22204-4012.

—. 1995. A Study of The Modern Language Aptitude Test for Predicting Learning Success and Advising Students. [Article submitted for publication] Available from the author at: Foreign Service Institute, US Department of State, 1501 S Randolph St, Arlington VA 22204-4012.

Ellis, Rod. 1986. *Understanding Second Language Acquisition*. Oxford: OUP.

English Language Teaching Development Unit with SKF Group. 1976. *Stages of Attainment Scale and Test Battery*. Oxford: OUP.

Erb, Karen; Harris, Martha; Heacock, Carolyn. 1994. Developing a Reading Proficiency Exam. TESOL 28th Annual Convention. Baltimore. Mar 8-12.

Evans, Helen. 1994 Jun. Modern Language Team Leader/subject Area Manager (English B). International Baccalaureate Organisation Curriculum & Assessment Office, Cardiff UK. [Personal communication to Bryan J Robinson].

Facultad de Traducción e Interpretación. 1993. *Acta de las pruebas específicas de aptitud*. Granada, Spain: University of Granda. Available from the Faculty Secretariat.

—. 1995. *Actas de Traducción General A-B (Inglés) 1994-95*. Granada, Spain: University of Granda. Available from the Faculty Secretariat.

—. 1995. *Actas De Traducción General B-A (Inglés) 1994-95*. Granada, Spain: University of Granda. Available from the Faculty Secretariat.

Fox WO. 1997(a) Mar 21 [E-mail message to Bryan J Robinson]. Available from the addressee at: Dept of Translation and Interpreting, University of Granada, Spain.

Fox WO. 1997(b). [Unpublished manuscript]. Available from the author at: Faculty of Translation and Interpreting, Autonomous University of Barcelona, Spain.

Fraser, Julia. 1996. Mapping the Process of Translation. *Meta* 41(1):84-96.

Gardner, WC; Lambert, WE. 1972. *Attitudes and Motivation in Second Language Learning*. Rowley, MA: Newbury House.

—. 1965. Language Aptitude, Intelligence, and Second Language Achievement. *Journal of Educational Psychology* 56:191-99.

Gillespie, Stuart, editor. 1992. *Translation and Literature*. 1. Edinburgh: Edinburgh University Press.

Grabe, W. 1991. Current Developments in Second Language Reading Research. *Tesol Quarterly* 25(3):431-59.

Grellet, Françoise. 1981. *Developing Reading Skills*. Cambridge: CUP.

Harris, Brian, compiler. 1997. *Translation and Interpreting Schools*. Amsterdam & Philadelphia: John Benjamins Publishing Co.

Harris, Michael; McCann, Paul. 1994. *Assessment*. Oxford: Heinemann.

Hartley J. 1994. *Designing Instructional Text*. 3rd ed. London: Kogan Page Ltd.

Hatim B, Ian Mason. 1990. *Discourse and the Translator*. London and New York: Longman Ltd.

Heaton, J. B. *Classroom Testing*. 1990. London & New York: Longman Ltd.

—. 1988. *Writing English Language Tests*. Harlow, Essex: Longman Group UK Ltd.

Henning, Grant. 1987. *A Guide to Language Testing. Development Evaluation Research*. Boston, Mass.: Heinle & Heinle Publishers.

Henning Pedersen, Niels; Picht, Heribert; Pugaard, Hanne. 1988. *Seminario Sobre Diseño y Desarrollo Curricular*. Granada: ICE de la Universidad de Granada.

Hess, Natalie. 1991. *Headstarts*. Harlow, Essex: Longman Group UK Ltd.

Holme, Randal. 1991. *Talking Texts*. Harlow, Essex: Longman Publishing UK Ltd.

House, ER. 1980. *Evaluating With Validity*. Beverly Hills, California: Sage.

Hughes, Arthur; Porter, D, editors. 1983. *Current Developments in Language Testing*. Londond: Academic Press.

Iser, Wolfgang. 1974. *The Implied Reader*. Baltimore & London: John Hopkins University Press.

Jolly, D. 1978. The Establishment of a Self-Access Scheme for Intensive Reading. Goethe Institute, British Council Colloquium on Reading. Paris. Oct.

Kelly, Dorothy A. 1989. International Course Committee Applied Languages Europe. [In-house working document for the purposes of the Applied Languages Europe programme. 2nd ed. Unpublished.] Available from Dorothy A. Kelly at: Dept of Translation and Interpreting, University of Granada, Spain.

Kendall, M. 1975. *Rank Correlation Methods*. 4th ed. London: Charles Griffin & Co Ltd.

Krashen, SD, Long, MA; Scarcella, RC. 1979. Age, Rate and Eventual Attainment in Second Language Acquisition. *TESOL Quarterly* 13:573-82.

Lannon, John M. 1994. *Technical Writing*. New York: HarperCollins College Publishers.

Larson-Freeman, Diane; Long, Michael H. 1991. *An Introduction to Second Language Acquisition Research*. Harlow, Essex: Longman Group UK Ltd.

Laycock, Liz. Date unknown. Testing Reading ... an Investigation. [Photocopied article. Source unknown]. Available from Bryan Robinson at: Dept of Translation and Interpreting, University of Granada, Spain.

Lörscher, Wolgang. 1986. Linguistic Aspects of Translation Processes: Towards an Analysis of Translation Performance. In House Juliane; Blum-Kulka, Shoshona, editors. *Interlingual and Intercultural Communication*. Tübingen: Narr:277-92.

—. 1991. *Translation Performance, Translation Process, and Translation Strategies. A Psycholinguistic Investigation*. Tübingen: Narr.

—. 1992(a). Process-Oreinted Research into Translation and Implications for Translation Teaching. *Traduction, Terminologie, Rédaction (TTR)* 5(1):5-61.

—. 1992(b). Translation Process Analysis. *Proceedings of the Fourth Scandinavian Symposium on Translation Theory: Translation and Knowledge*. 195-211.

Lumley, T. 1993. The Notion of Subskills in Reading Comprehension Tests: an EAP Example. *Language Testing* 10:211-34.

Lumley, T; Alderson, J Charles. 1995. Responses and Replies: The Notion of Subskills in Reading Comprehension Tests: an EAP Example. *Language Testing* 12(1):121-30.

Madsen, HS. 1983. *Techniques in Testing*. New York & Oxford: OUP, 1983.

Mahn, Gabriela. 1987. Symposium: Foreign Language Proficiency Criteria in Translation. In: Rose, Marilyn Gaddis, editor. *Translation Excellence: Assessment, Achievement, Maintenance*. New York: State University of New York at Binghamton.

—. 1989. Standards and Evaluation in Translator Training. In: Krawutschke, Peter W., editor. *Translator and Interpreter Training and Foreign Language Pedagogy*. New York: State University of New York at Binghamton:100-08.

McLaughlin, B. 1980. Theory and Research in Second Language Learning: an Emerging Paradigm. *Language Learning* 30(2):331-50.

Monikowski, Christine. 1994. Developing and Administering a Cloze Test in American Sign Language. 16th Annual Language Testing Research Colloquium. Mar 5-7. Washington DC.

Moore, John, et al. 1979. *Discovering Discourse*. Oxford: OUP.

Moroney, MJ. 1953. *Facts From Figures*. London: Penguin Books.

Muir, Kenneth. 1992. Translating Golden Age Plays: A Reconsideration. *Translation and Literature* 1:104-11.

Muñiz Fernández, José. 1990. *Teoría De Respuesta a Los Ítems*. Madrid: Ediciones Pirámide S.A.

Nilski, Thérèse. 1967. Translators and Interpreters: Siblings or a Breed Apart? *Meta* 12(2):45-49.

Oller, JW, editor. 1983. *Issues in Language Testing Research*. Rowley, Mass.: Newbury House.

Pérez Basanta, Carmen. 1993. [Handwritten notes and photocopies from a course on Testing.] Available from Bryan Robinson at: Dept of Translation and Interpreting, University of Granada, Spain.

Perkins, Kyle. 1994. Predicting Item Difficulty in a Reading Comprehension Test With an Artificial Neural Network . 16th Annual Language Testing Research Colloquium. Mar 5-7. Washington DC.

Perren, GE. 1977. *Foreign Language Testing: Special Bibliography*. London: Centre for Information on Language Teaching and Research.

Phillips, E; Pugh, DS. 1994. *How to Get a Ph.D. A Handbook for Students and Their Supervisors*. Buckingham: Open University Press.

Plackett, Elizabeth. Date unknown. How I Stopped Testing and Learnt to Live Without It. [Photocopied article. Source unknown]. Available from Bryan Robinson at: Dept of Translation and Interpreting, University of Granada, Spain.

Pollard, John. 1994. Paper on Proficiency Testing. *Testing Newsletter* . Whitstable, Kent: IATEFL SIG On Testing:36-56.

Pollitt, Alastair. 1991. Giving Students a Sporting Chance: Assessment by Counting and by Judging.:46-59.

Rainey, Brian E. 1988. Thoughts on the 'Language Crisis' and the Teaching of Translation. In: Hammond, Deanna L, editor. *Languages at Crossroads, Proceedings of the 29th Annual Conference of the American Translators Association*. Medford, NJ: Learned Information Inc.:291-96.

Reiss, Katerina. 1992. *Modelo de factores que influyen en la traducción*. [Photocopied handout distributed at a public lecture givene in the University of Granada.] Available from Bryan Robinson at: Dept of Translation and Interpreting, University of Granada, Spain.

Robinson, Bryan J, reviewer. 1999. [Review of Bachman, Lyle F.; Palmer, Adrian S. 1996. *Language Testing in Practice*. London: OUP.]. In: *Modern English Teacher* 8(1):83-85.

—. 1993(a). *Paper One Text-handling. Higher level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1993(b). *Paper Two Written production. Higher level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1993(c). *Paper One Text-handling. Standard level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1993(d). *Paper Two Written production. Standard level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1994(a). *Paper One Text-handling. Higher level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1994(b). *Paper Two Written production. Higher level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1994(c). *Paper One Text-handling. Standard level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1994(d). *Paper Two Written production. Standard level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1995(a). *Paper One Text-handling. Higher level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1995(b). *Paper Two Written production. Higher level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1995(c). *Paper One Text-handling. Standard level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1995(d). *Paper Two Written production. Standard level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1996(a). *Paper One Text-handling. Higher level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1996(b). *Paper Two Written production. Higher level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1996(c). *Paper One Text-handling. Standard level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1996(d). *Paper Two Written production. Standard level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1997(a). *Paper One Text-handling. Higher level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1997(b). *Paper Two Written production. Higher level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1997(c). *Paper One Text-handling. Standard level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1997(d). *Paper Two Written production. Standard level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1998(a). *Paper One Text-handling. Higher level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1998(b). *Paper Two Written production. Higher level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1998(c). *Paper One Text-handling. Standard level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1998(d). *Paper Two Written production. Standard level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1999(a). *Paper One Text-handling. Higher level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1999(b). *Paper Two Written production. Higher level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1999(c). *Paper One Text-handling. Standard level*. Cardiff & Geneva: International Baccalaureate Organisation.

—. 1999(d). *Paper Two Written production. Standard level*. Cardiff & Geneva: International Baccalaureate Organisation.

Rowntree, Derek. 1981. *Statistics Without Tears*. London: Penguin Books.

Rudska, B., et al. 1981. *The Words You Need*. London: Macmillan Press.

Schmitt, Norbert. 1995. An Examinationof the Behaviour of Four Vocabulary Tests. In: Allan, Dave, editor. *Entry Points. Papers From a Aymposium of Research, Testing and Young Learners Special Interest Groups*. Whitstable, Kent: IATEFL SIG on Testing.

Seliger, Herbert W; Shohamy, Elana. 1989. *Second Language Research Methods*. Oxford English. Oxford: OUP.

Siegel, S; Castellan, NJ Jr. 1988. *Nonparametric Statistics for the Behavioural Sciences*. 2nd ed.: McGraw-Hill.

Silver, Mick. 1997. *Business Statistics*. 2nd ed. Maidenhead, Berks: McGraw-Hill.

Skehan, Peter. 1986. The Role of Foreign Language Aptitude in a Model of School Learning. *Language Testing* .3:188-221.

—. 1988. Language Testing: Survey Article, Part I. *Language Teaching Abstracts* 21(4):211-21.

—. 1989. Language Testing: Survey Article, Part II. *Language Teaching Abstracts* 22(1):1-13.

Snell-Hornby, Mary. 1988. *Translation Studies. An Integrated Approach.* Amsterdam: John Benjamins Publishing Co.

—. 1994. An Overview of Doctoral Research in Vienna. [Notes taken at a lecture given in the University of Granada.] Available from Bryan Robinson at: Dept of Translation and Interpreting, University of Granada, Spain.

Spearman, C. 1904. General Intelligence, Objectively Determined and Measured. *American Journal of Psychology* 15:201-93.

—. 1910: Correlation Calculated From Faulty Data. *British Journal of Psychology* 3:271-95.

—. 1927. *The Abilities of Man: Their Nature and Measurement.* New York: AMS Publishers.

Spolsky, Bernard. 1995. *Measured Words*. Oxford: OUP.

Stanley, JC. 1971. *Educational Measurement.* (ed) Thorndike, RL. 2nd ed. 356-442.

Sydes, M; Hartley, James. 1997. A Thorn in the Flesch: Observations on the Unreliability of Computer-Based Readability Formulae. *British Journal of Educational Technology* 28(2):143-45.

Szuki, Atsuko. 1988. Aptitudes of Translators and Interpreters. *Meta* 33(1):108-14.

Taylor, W. L. 1953. Cloze Procedure: A New Tool for Measuring Readability. *Journalism Quarterly* 30:415-33.

Thorndike, RL. 1978. *Correlational Procedures for Research.* New York: Gardner Press.

Tirkkonen-Condit, Sonja, editor. 1991. *Empirical Research in Translation and Intercultural Studies: Selected Papers of the TRANSIF Seminar, Savonlinna,* Language in Performance 5. Tübingen: Gunter Narr Verlag Tübingen.

Traugott, Elizabth C; Marie Louise Pratt. 1980. *Linguistics for Students of Literature*. New York: Harcourt Brace, Jovanich.

Verschoor, Alfred. 1997 Feb 18. [E-mail to Bryan Robinson]. Available from Bryan J Robinson at: Dept of Translation and Interpreting, University of Granada, Spain.

Vigneault, R. 1998. Subject Area Manager (English B). International Baccalaureate Organisation Curriculum & Assessment Office, Cardiff UK. [Personal communication to Bryan J Robinson].

Wall, Dianne, Caroline Clapham; J Charles Alderson. 1991. Validating Tests in Difficult Circumstances. In: Alderson, J. Charles; North, Brian, editors. *Language Testing in the 1990s: The Communicative Legacy*. London: Modern English Publications and The British Council:209-25.

Weir, Cyril J. 1983. *Identifying the Language Needs of Overseas Students in Tertiary Education in the United Kingdom*. Microfilm Unpublished PhD thesis. Institute of Education, University of London.

—. 1988. *Communicative Language Testing*. Language Teaching Methodology: Prentice Hall International.

—. 1993. *Understanding and Developing Language Tests*. Prentice Hall International English Language Teaching. London: Prentice-Hall.

West, David. 1992. Translating the *Aeneid*. *Translation and Literature* 1:97-103.

Wood, Robert. 1991. *Assessment and Testing: A Survey of Research*. Cambridge: CUP.

Woods, Anthony J; Fletcher, Paul; Hughes, Arthur. 1986. *Statistics in Language Studies*. Cambridge Textbooks in Linguistics. Cambridge: CUP.

Wright, BR; Stone, MH. 1979. *Best Test Design*. Chicago, Ill: MESA Press.

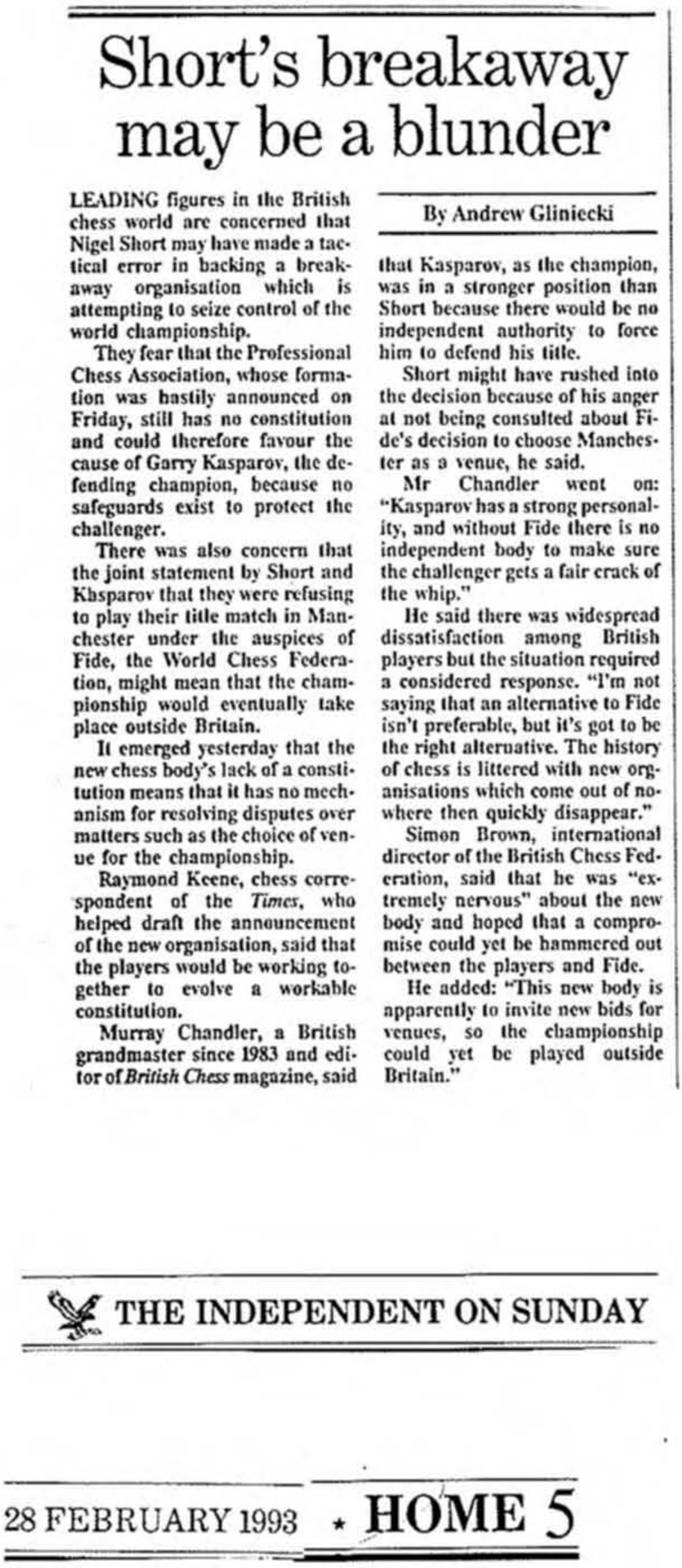
Wu Yi'an. 1998. What Do Tests of Listening Comprehension Test? - A Retrospection Study of EFL Test-Takers Performing a Multiple Choice Task. *Language Testing* 15(1):21-44.

ERRATA

Page	*Line*	*Errata*	*Correction*
44	1	These are crude labels that we attach to *r* values	These are crude labels that we attach to *r* values (Cohen and Mannion (1989:168-69)
100	5	...analyzed on individually...	...analyzed individually...
155	19	Easy 0.80	Easy ≥0.80
	24	DI 0.30	DI ≥0.30
188	37	Our target score for all items was 0.30...	Our target score for all items was ≥0.30...
197	3	...the combination of skills 22.1— ...	...the combination of skills 32.1— ...
225	23	Data omitted	Variance:4.37;5.21; 6.46
248	28	Data omitted	Variance 4.99;5.00;5.49
257	2	*The Guardian* 1993.	*The Guardian* 9 June1993.
259	14	The target range for these items was 0.80...	The target range for these items was ≥0.80...
276	8	Data omitted	Variance: 5.53
306	12	Data omitted	Variance: 1.88
	23	Data omitted	Variance: 8.42
	32	Data omitted	Variance: 4.00
307	9	Data omitted	Variance: 10.33
336	13	Data omitted	Variance: 1.94
	23	Data omitted	Variance: 7.48
	32	Data omitted	Variance: 3.34
337	9	Data omitted	Variance: 10.35
342	14	SPEARMAN-BROWN PROPHECY FORMULA (18—9)	SPEARMAN-BROWN PROPHECY FORMULA (18—16)
353	27	Table 18—5	Table 18—10
354	2	Table 18—8	Table 18—11
	11	Table 18—8	Table 18—11
	16	Table 18—9	Table 18—12
	25	Table 18—8	Table 18—11
355	17	Table 18—10	Table 18—13
356	1	Table 18—10	Table 18—13
356	22	Table 18—11	Table 18—14
357	21	Table 18—12	Table 18—15
358	26	Table 18—9	Table 18—16
359	31-32	Tables 18—10 and 18—11	Tables 18—17 and 18—18
360	1	Table 18—10	Table 18—17
	7	Table 18—11	Table 18—18
	16	Tables 18—12 and 18—13	Tables 18—19 and 18—20
	17	Table 18—12	Table 18—19
361	1	Table 18—13	Table 18—20
362	25	...our target for discrimination of 0.30...	...our target for discrimination of ≥0.30...

Appendices

Sub-test in01.1



Short's breakaway may be a blunder

By Andrew Gliniecki

LEADING figures in the British chess world are concerned that Nigel Short may have made a tactical error in backing a breakaway organisation which is attempting to seize control of the world championship.

They fear that the Professional Chess Association, whose formation was hastily announced on Friday, still has no constitution and could therefore favour the cause of Garry Kasparov, the defending champion, because no safeguards exist to protect the challenger.

There was also concern that the joint statement by Short and Kasparov that they were refusing to play their title match in Manchester under the auspices of Fide, the World Chess Federation, might mean that the championship would eventually take place outside Britain.

It emerged yesterday that the new chess body's lack of a constitution means that it has no mechanism for resolving disputes over matters such as the choice of venue for the championship.

Raymond Keene, chess correspondent of the *Times*, who helped draft the announcement of the new organisation, said that the players would be working together to evolve a workable constitution.

Murray Chandler, a British grandmaster since 1983 and editor of *British Chess* magazine, said that Kasparov, as the champion, was in a stronger position than Short because there would be no independent authority to force him to defend his title.

Short might have rushed into the decision because of his anger at not being consulted about Fide's decision to choose Manchester as a venue, he said.

Mr Chandler went on: "Kasparov has a strong personality, and without Fide there is no independent body to make sure the challenger gets a fair crack of the whip."

He said there was widespread dissatisfaction among British players but the situation required a considered response. "I'm not saying that an alternative to Fide isn't preferable, but it's got to be the right alternative. The history of chess is littered with new organisations which come out of nowhere then quickly disappear."

Simon Brown, international director of the British Chess Federation, said that he was "extremely nervous" about the new body and hoped that a compromise could yet be hammered out between the players and Fide.

He added: "This new body is apparently to invite new bids for venues, so the championship could yet be played outside Britain."

THE INDEPENDENT ON SUNDAY

28 FEBRUARY 1993 ★ HOME 5

Surname(s): ____________________ Name: ____________

Short's breakaway may be a blunder
Published in The Independent on Sunday, 28 Feb 1993, p. 5

Scan the text and identify which paragraphs are summarised by the five sentences below. (12b)

1. Short's motives 7
2. A joint statement 3
3. Help from a journalist 5
4. Kasparov's personality 8
5. Other players' reactions 9

True/false/questionable (10 & 8)

6. Manchester may be the venue chosen to host the world title match Q

7. The new organisation is called *Fide* F

8. Short is the favourite to win the title F

Complete the following sentences. (3)

9. The venue of the title match ____________

10. In the newly established chess organisation Short ____________

(5) 11. The breakaway organisation (para 1)

A has been established by Raymond Keene
B has chosen a new venue for the match
C is supported by Garry Kasparov
[✓ D] was set up by Kasparov and Short
E is called the British Chess Federation

12. The challenger (para 8) is

A Murray Chandler
B Garry Kasparov
C Simon Brown
[✓ D] Nigel Short
E None of these

Choose the alternative you think gives the most suitable meaning. (2)

13. "a fair crack of the whip" (para 8) means

A a beating
[✓ B] an equal opportunity
C an advantage
D an argument
E a noisy response

14. "hammered out" (para 10) means

A simply rejected
B attacked violently
[✓ C] agreed upon
D imposed by others
E discussed calmly

(6) 15. The sentence "Kasparov has a strong personality ... of the whip" (para 8) is intended as

A an assessment
[✓ B] a warning
C a new insight
D a joke
E a forlorn hope

16. The sentence "I'm not saying that ... then quickly disappear" (para 9) serves as

[✓ A] a balanced qualification
B a new alternative
C the definitive response
D the last chance
E a tried and tested response

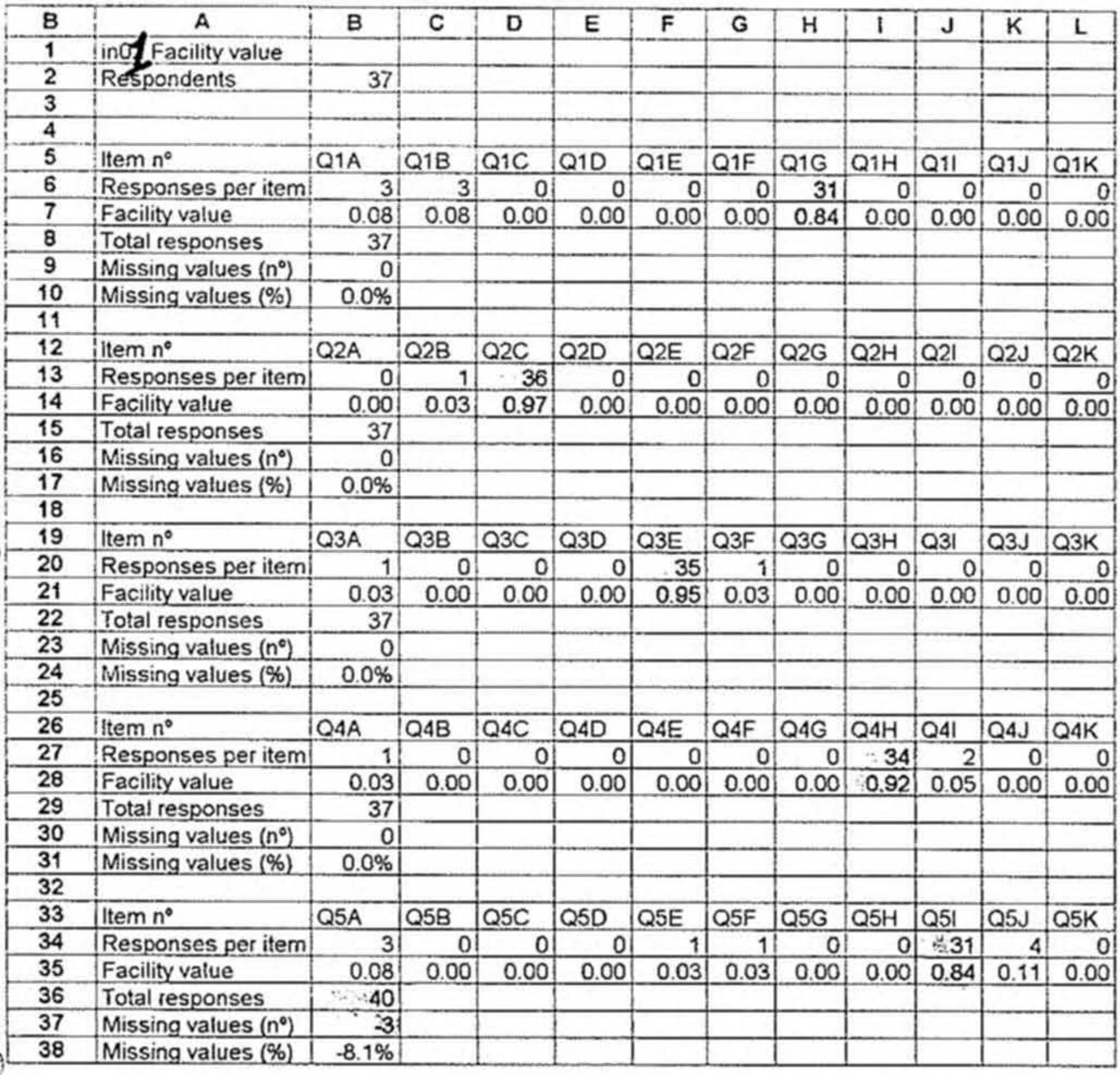
17. Fill in the boxes below in such a way as to summarise the main content of the text. You don't need to use full sentences - you will be marked for the information you give, and not for the way it is expressed. (11)

Event:	
Cause(s):	1. 2.
Consequence(s):	1. 2. 3. 4.

18. The following paragraph is a summary of the text, but it is too long. Delete those words and/or phrases which you think are unnecessary and reduce it to 50 words. (13c)

A dispute between the world champion and the challenger has led to the establishment of a new world-wide chess organisation which hopes to be able to take control of the sport. However, there is still a lot of concern about this as the new organisation has yet to agree on its constitution, and this lays it open to abuse by its founders. Whatever the outcome, the next title match is likely to be played at a venue outside of Britain.

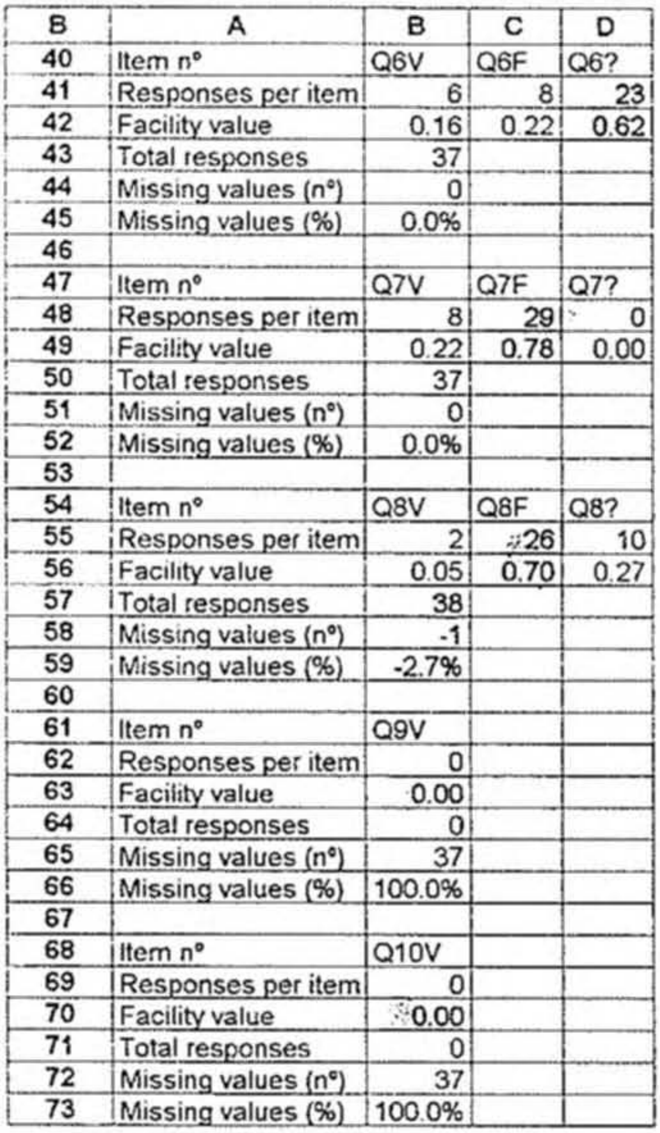
in01 - Facility value - eutistat.wb1 - ~~Hoja B - 13/3/96~~



B	A	B	C	D	E	F	G	H	I	J	K	L
1	in01 Facility value											
2	Respondents	37										
3												
4												
5	Item n°	Q1A	Q1B	Q1C	Q1D	Q1E	Q1F	Q1G	Q1H	Q1I	Q1J	Q1K
6	Responses per item	3	3	0	0	0	0	31	0	0	0	0
7	Facility value	0.08	0.08	0.00	0.00	0.00	0.00	0.84	0.00	0.00	0.00	0.00
8	Total responses	37										
9	Missing values (n°)	0										
10	Missing values (%)	0.0%										
11												
12	Item n°	Q2A	Q2B	Q2C	Q2D	Q2E	Q2F	Q2G	Q2H	Q2I	Q2J	Q2K
13	Responses per item	0	1	36	0	0	0	0	0	0	0	0
14	Facility value	0.00	0.03	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	Total responses	37										
16	Missing values (n°)	0										
17	Missing values (%)	0.0%										
18												
19	Item n°	Q3A	Q3B	Q3C	Q3D	Q3E	Q3F	Q3G	Q3H	Q3I	Q3J	Q3K
20	Responses per item	1	0	0	0	35	1	0	0	0	0	0
21	Facility value	0.03	0.00	0.00	0.00	0.95	0.03	0.00	0.00	0.00	0.00	0.00
22	Total responses	37										
23	Missing values (n°)	0										
24	Missing values (%)	0.0%										
25												
26	Item n°	Q4A	Q4B	Q4C	Q4D	Q4E	Q4F	Q4G	Q4H	Q4I	Q4J	Q4K
27	Responses per item	1	0	0	0	0	0	0	34	2	0	0
28	Facility value	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.05	0.00	0.00
29	Total responses	37										
30	Missing values (n°)	0										
31	Missing values (%)	0.0%										
32												
33	Item n°	Q5A	Q5B	Q5C	Q5D	Q5E	Q5F	Q5G	Q5H	Q5I	Q5J	Q5K
34	Responses per item	3	0	0	0	1	1	0	0	31	4	0
35	Facility value	0.08	0.00	0.00	0.00	0.03	0.03	0.00	0.00	0.84	0.11	0.00
36	Total responses	40										
37	Missing values (n°)	-3										
38	Missing values (%)	-8.1%										

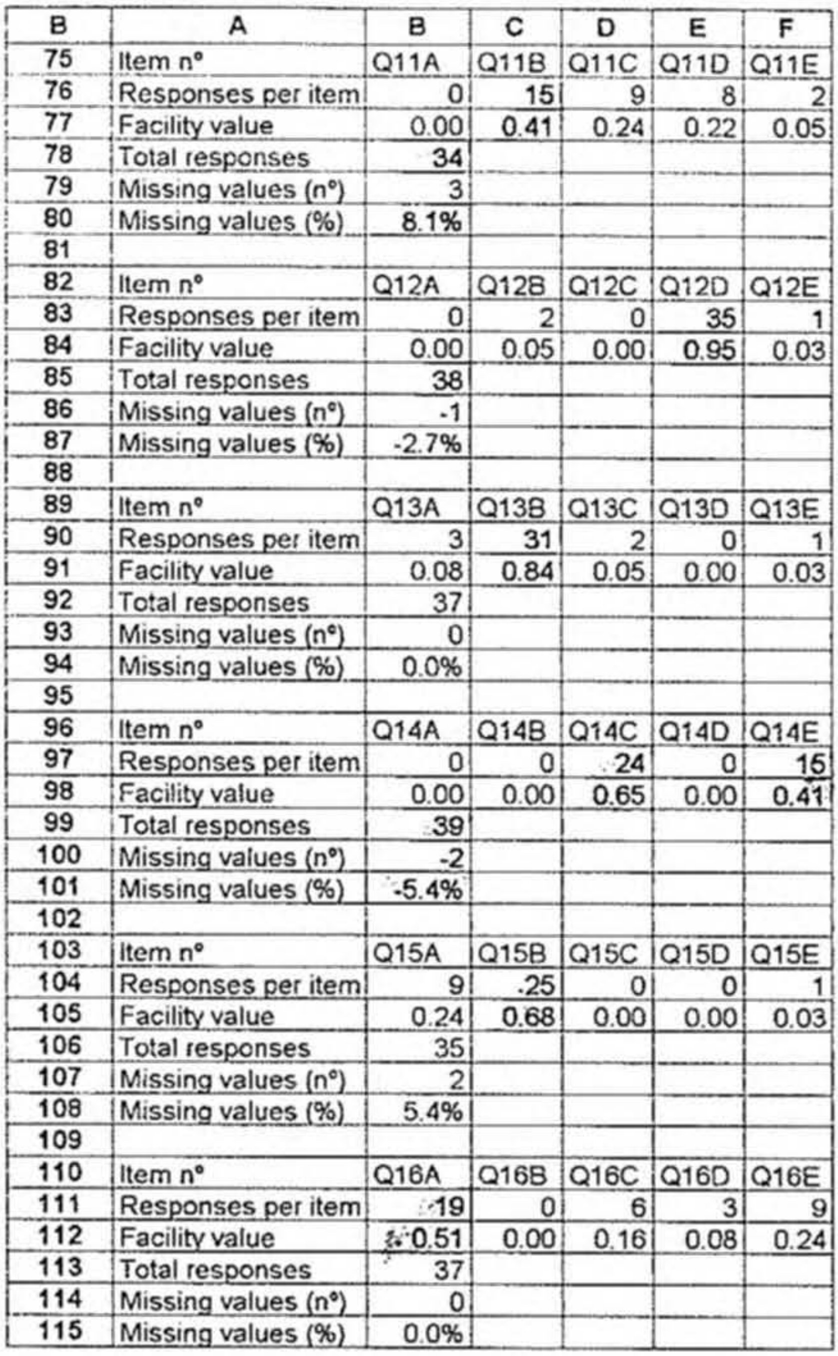
Discount N° 5

in01 - Facility value - eulistat.wb1 - Hoja B - 13/3/96



B	A	B	C	D
40	Item nº	Q6V	Q6F	Q6?
41	Responses per item	6	8	23
42	Facility value	0.16	0.22	0.62
43	Total responses	37		
44	Missing values (nº)	0		
45	Missing values (%)	0.0%		
46				
47	Item nº	Q7V	Q7F	Q7?
48	Responses per item	8	29	0
49	Facility value	0.22	0.78	0.00
50	Total responses	37		
51	Missing values (nº)	0		
52	Missing values (%)	0.0%		
53				
54	Item nº	Q8V	Q8F	Q8?
55	Responses per item	2	26	10
56	Facility value	0.05	0.70	0.27
57	Total responses	38		
58	Missing values (nº)	-1		
59	Missing values (%)	-2.7%		
60				
61	Item nº	Q9V		
62	Responses per item	0		
63	Facility value	0.00		
64	Total responses	0		
65	Missing values (nº)	37		
66	Missing values (%)	100.0%		
67				
68	Item nº	Q10V		
69	Responses per item	0		
70	Facility value	0.00		
71	Total responses	0		
72	Missing values (nº)	37		
73	Missing values (%)	100.0%		

in01 - Facility value - eutistat.wb1 - Hoja B - 13/3/96



B	A	B	C	D	E	F
75	Item n°	Q11A	Q11B	Q11C	Q11D	Q11E
76	Responses per item	0	15	9	8	2
77	Facility value	0.00	0.41	0.24	0.22	0.05
78	Total responses	34				
79	Missing values (n°)	3				
80	Missing values (%)	8.1%				
81						
82	Item n°	Q12A	Q12B	Q12C	Q12D	Q12E
83	Responses per item	0	2	0	35	1
84	Facility value	0.00	0.05	0.00	0.95	0.03
85	Total responses	38				
86	Missing values (n°)	-1				
87	Missing values (%)	-2.7%				
88						
89	Item n°	Q13A	Q13B	Q13C	Q13D	Q13E
90	Responses per item	3	31	2	0	1
91	Facility value	0.08	0.84	0.05	0.00	0.03
92	Total responses	37				
93	Missing values (n°)	0				
94	Missing values (%)	0.0%				
95						
96	Item n°	Q14A	Q14B	Q14C	Q14D	Q14E
97	Responses per item	0	0	24	0	15
98	Facility value	0.00	0.00	0.65	0.00	0.41
99	Total responses	39				
100	Missing values (n°)	-2				
101	Missing values (%)	-5.4%				
102						
103	Item n°	Q15A	Q15B	Q15C	Q15D	Q15E
104	Responses per item	9	25	0	0	1
105	Facility value	0.24	0.68	0.00	0.00	0.03
106	Total responses	35				
107	Missing values (n°)	2				
108	Missing values (%)	5.4%				
109						
110	Item n°	Q16A	Q16B	Q16C	Q16D	Q16E
111	Responses per item	19	0	6	3	9
112	Facility value	0.51	0.00	0.16	0.08	0.24
113	Total responses	37				
114	Missing values (n°)	0				
115	Missing values (%)	0.0%				

Poor item

Good item.

Test specification

Cohort A

Licenciatura en Traducción e Interpretación

PROPUESTA DE PRUEBA DE APTITUD

1 Especificaciones

1.1 Formato

La prueba constará de tres ejercicios obligatorios:

1º Prueba de habilidades lectivas en lengua A (Español)

2º Prueba de habilidades lectivas en lengua B (Alemán, francés o inglés)

3º Prueba de habilidades auditivas en lengua B

1.2 Puntuación

El objetivo del sistema de puntuación es el de adjudicar las plazas vacantes a los alumnos más adeptos de entre los que se presentan.

El primer ejercicio tendrá carácter eliminatorio. El segundo y el tercer ejercicio se evaluarán conjuntamente.

1.3 Descripción de los ejercicios

1 1er ejercicio

Los candidatos deberán responder a una serie de preguntas de elección múltiple sobre un texto periodístico de actualidad, de una extensión aproximada de seiscientas palabras, con el fin de determinar su capacidad de usar las siguientes habilidades lectivas:

1 Deducir el significado y uso de unidades léxicas mediante la comprensión de la morfología y el contexto

2 Comprender las relaciones establecidas entre los componentes de una frase

3 Comprender las relaciones establecidas entre distintos componentes del texto que reflejan los nexos gramaticales

4 Comprender las relaciones establecidas entre distintos componentes del texto mediante el reconocimiento de indicadores de discurso

5 Comprender la función comunicativa de la frase según se empleen, o no, indicadores de discurso

6 Comprender las relaciones conceptuales, p.ej. causa-efecto

7 Comprender ideas explícitas

8 Comprender ideas implícitas

9 Separar el contenido esencial de los no esenciales, distinguir la idea principal de los detalles que la ilustran

10 Transferir información desde un medio a otro, p-ej. desde un texto a un gráfico

11 Examinar el texto rápidamente para obtener información específica

12 Tomar apuntes. Reducir el texto mediante la eliminación del contenido redundante o irrelevante

2 2º ejercicio

Los candidatos deberán responder a una serie de preguntas de elección múltiple sobre un texto periodístico de actualidad - pero no de opinión - y de una extensión aproximada de cuatrocientas palabras, con el fin de determinar su capacidad de usar sus habilidades lectivas en la lengua B.

3 3er ejercicio

Los textos a utilizar serán monólogos grabados de medios de comunicación públicos - radio o televisión. Serán de interés general, sin ser de opinión, ni semi-especializados. El total de la grabación durará unos siete u ocho minutos. Los acentos reflejarán variedades de uso común a nivel internacional: p. ej. Inglés británico e inglés americano.

Se formularán las preguntas con el fin de averiguar la capacidad de los candidatos para usar sus habilidades auditivas.

2 <u>Tribunal</u>

El presidente del tribunal será el Director de la EUTI, o la persona en que él delegue la responsabilidad. Además del presidente, el tribunal estará compuesto por un secretario/coordinador de las pruebas de aptitud y nueve vocales, en representación de las lenguas B.

Sub-test es01.2

INFORMACION SOBRE EL EXAMEN

Este examen consta de dos partes que son pruebas de habilidades de lectura.

Antes de comenzar comprueba haber rellenado los datos personales en las hojas de respuestas, y haber entendido bien como señalar las respuestas correctas.

La primera prueba es un texto en lengua española; la segunda es un texto en alemán, francés, o inglés.

El tiempo total de duración del examen es de 2 horas. La distribución del tiempo se deja al arbitrio de cada uno, pero se sugiere emplear alrededor de una hora para cada parte.

Está prohibido salirse del examen sin entregar los papeles. En cualquier caso no se permitirá la salida durante los últimos diez minutos del examen.

No se admitirá el uso de ningún libro, papel o cualquier otro tipo de ayuda. Las personas que los utilicen se descalificarán automáticamente.

Se recomienda el uso de lápiz y goma de borrar para el examen. No es necesario el uso de bolígrafo.

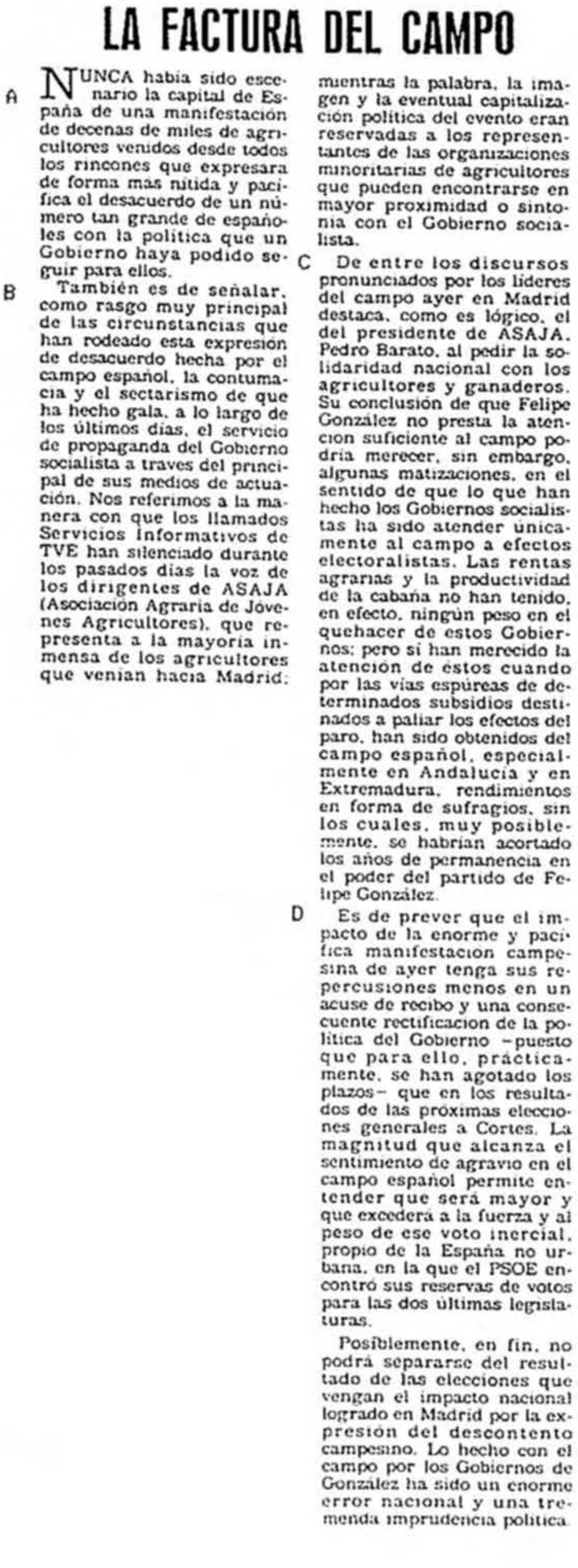
Se puede usar la hoja de preguntas como borrador, pero todas las respuestas deberán consignarse en las hojas de respuestas.

El formato de las preguntas es de elección múltiple. Las instrucciones son de la siguiente manera:

> Lee el texto y señala a cuál de los párrafos podría corresponder cada una de las siguientes frases como título. Subraya la letra en la hoja de respuestas.
>
> y
>
> V (Verdadero), F (Falso) o ? (el texto no informa sobre este punto), y subraya la letra o el símbolo correspondiente en la hoja de respuestas.
>
> y
>
> Elige la opción más apropiada y subraya la letra en la hoja de respuestas.

Si se comete algún error al rellenar la hoja de respuestas borra la respuesta errónea y subraya la correcta, o tacha con una "X" la respuesta errónea y subraya la correcta.

Al final del examen se entregarán todos los papeles: es decir, textos, hojas de preguntas, y hojas de respuestas.



LA FACTURA DEL CAMPO

A NUNCA habia sido escenario la capital de España de una manifestación de decenas de miles de agricultores venidos desde todos los rincones que expresara de forma mas nitida y pacífica el desacuerdo de un número tan grande de españoles con la politica que un Gobierno haya podido seguir para ellos.

B También es de señalar, como rasgo muy principal de las circunstancias que han rodeado esta expresión de desacuerdo hecha por el campo español, la contumacia y el sectarismo de que ha hecho gala, a lo largo de los últimos dias, el servicio de propaganda del Gobierno socialista a traves del principal de sus medios de actuación. Nos referimos a la manera con que los llamados Servicios Informativos de TVE han silenciado durante los pasados dias la voz de los dirigentes de ASAJA (Asociación Agraria de Jóvenes Agricultores), que representa a la mayoria inmensa de los agricultores que venian hacia Madrid; mientras la palabra, la imagen y la eventual capitalización politica del evento eran reservadas a los representantes de las organizaciones minoritarias de agricultores que pueden encontrarse en mayor proximidad o sintonia con el Gobierno socialista.

C De entre los discursos pronunciados por los líderes del campo ayer en Madrid destaca, como es lógico, el del presidente de ASAJA, Pedro Barato, al pedir la solidaridad nacional con los agricultores y ganaderos. Su conclusión de que Felipe González no presta la atencion suficiente al campo podria merecer, sin embargo, algunas matizaciones, en el sentido de que lo que han hecho los Gobiernos socialistas ha sido atender únicamente al campo a efectos electoralistas. Las rentas agrarias y la productividad de la cabaña no han tenido, en efecto, ningún peso en el quehacer de estos Gobiernos; pero sí han merecido la atención de éstos cuando por las vías espúreas de determinados subsidios destinados a paliar los efectos del paro, han sido obtenidos del campo español, especialmente en Andalucía y en Extremadura, rendimientos en forma de sufragios, sin los cuales, muy posiblemente, se habrían acortado los años de permanencia en el poder del partido de Felipe González.

D Es de prever que el impacto de la enorme y pacífica manifestacion campesina de ayer tenga sus repercusiones menos en un acuse de recibo y una consecuente rectificacion de la política del Gobierno –puesto que para ello, prácticamente, se han agotado los plazos– que en los resultados de las próximas elecciones generales a Cortes. La magnitud que alcanza el sentimiento de agravio en el campo español permite entender que será mayor y que excederá a la fuerza y al peso de ese voto inercial, propio de la España no urbana, en la que el PSOE encontró sus reservas de votos para las dos últimas legislaturas.

Posiblemente, en fin, no podrá separarse del resultado de las elecciones que vengan el impacto nacional logrado en Madrid por la expresión del descontento campesino. Lo hecho con el campo por los Gobiernos de González ha sido un enorme error nacional y una tremenda imprudencia politica.

Español - Prueba de habilidades de lectura - Texto Nº 1

1er APELLIDO: □□□□□□□□□□□□□□□□□□□□□□□□□□

2º APELLIDO: □□□□□□□□□□□□□□□□□□□□□□□□□□

NOMBRE: □□□□□□□□□□□□□□□□□□□

DNI: □□□□□□□□□ [SCORE: □□]

Subraya la letra o símbolo que corresponda.

1.	A	B	C	D	E
2.	A	B	C	D	E
3.	A	B	C	D	E
4.	A	B	C	D	E
5.	V	F	?		
6.	V	F	?		
7.	V	F	?		
8.	V	F	?		
9.	V	F	?		
10.	V	F	?		
11.	A	B	C	D	E
12.	A	B	C	D	E
13.	A	B	C	D	E
14.	A	B	C	D	E
15.	A	B	C	D	E
16.	A	B	C	D	E
17.	A	B	C	D	E
18.	A	B	C	D	E
19.	A	B	C	D	E
20.	A	B	C	D	E

Español - Prueba de habilidades de lectura - Texto Nº 1

"La factura del campo"

Lee el texto y señala a cuál de los párrafos podría corresponder cada una de las siguientes frases como título. Subraya la letra en la hoja de respuestas.

1. "El acontecimiento".

2. "La actitud del gobierno hacia el campo".

3. "El papel del Telediario".

4. "Las próximas elecciones".

V (Verdadero), F (Falso) o ? (el texto no informa sobre este punto), y subraya la letra o el símbolo correspondiente en la hoja de respuestas.

5. Es muy significativo que esta manifestación haya tenido lugar en Madrid.

6. Pedro Barato opina que los socialistas perderán las próximas elecciones.

7. Los servicios informativos de TVE siempre se muestran parciales.

8. Si no fuera por el voto del campo el PSOE nunca habría llegado a gobernar.

9. En opinión del autor de este artículo el campo seguirá apoyando al Gobierno a pesar de lo hecho.

10. Las consecuencias más importantes de la manifestación se verán especialmente en Andalucía y Extremadura.

Elige la opción más apropiada y subraya la letra en la hoja de respuestas.

11. En opinión del autor de este texto, la conclusión de Pedro Barato es ...

A correcta.

B incompleta.

C electoralista.

D suficiente.

E superflua.

12. Como resultado de la presentación de esta noticia por parte de TVE __________ se ha(n) visto perjudicado(s).

A ... ASAJA ...

B ... organizaciones minoritarias de agricultores ...

C ... ningún colectivo ...

D ... todos ...

E ... el PSOE ...

13. Una rectificación por parte del Gobierno ...

A influiría en las elecciones.

B cambiaría la política de TVE.

C sería muy bien recibido.

D tendría efectos retroactivos.

E llegaría demasiado tarde.

14. ¿Qué es lo que el Gobierno ha obtenido?

A Subsidios.

B Votos.

C Beneficios.

D Rentas agrarias.

E Apoyos.

15. ¿Qué será mayor?

A La mayoría del nuevo gobierno.

B El sentimiento en contra del PSOE.

C La inercia del campo.

D La próxima manifestación.

E La respuesta urbana.

16. Ninguna manifestación anterior ha ...

A unido las organizaciones del campo.

B obligado al Gobierno a reaccionar.

C recibido tanto apoyo popular.

D destacado la proximidad de Gobierno y agricultores.

E demostrado tan claramente la actitud en contra del Gobierno.

17. La frase "la contumacia y el sectarismo de que ha hecho gala" en el párrafo B, quieren decir que TVE se ha mostrado ...

A neutral.
B irresponsable.
C partidaria del gobierno.
D inconsistente.
E influída por un sector.

18. La frase "la productividad de la cabaña" en el párrafo C refiere a ...

A los impuestos pagados por los agricultores.
B los ingresos que reciben al vender su ganado.
C el volumen de carne producido en España.
D los beneficios del sector ganadero.
E la cuantía de cereales de la cosecha anual.

19. La frase "las vías espúreas" en el párrafo C, son ...

A rutas peligrosas.
B métodos ilegítimos.
C la red viaria.
D un sistema burocrático.
E un planteamiento político.

20. La frase "en un acuse de recibo" en el párrafo D, significa ...

A un formulario oficial.
B una respuesta por escrito.
C una reacción protocolaria.
D un reconocimiento formal.
E una acusación jurídica.

(1

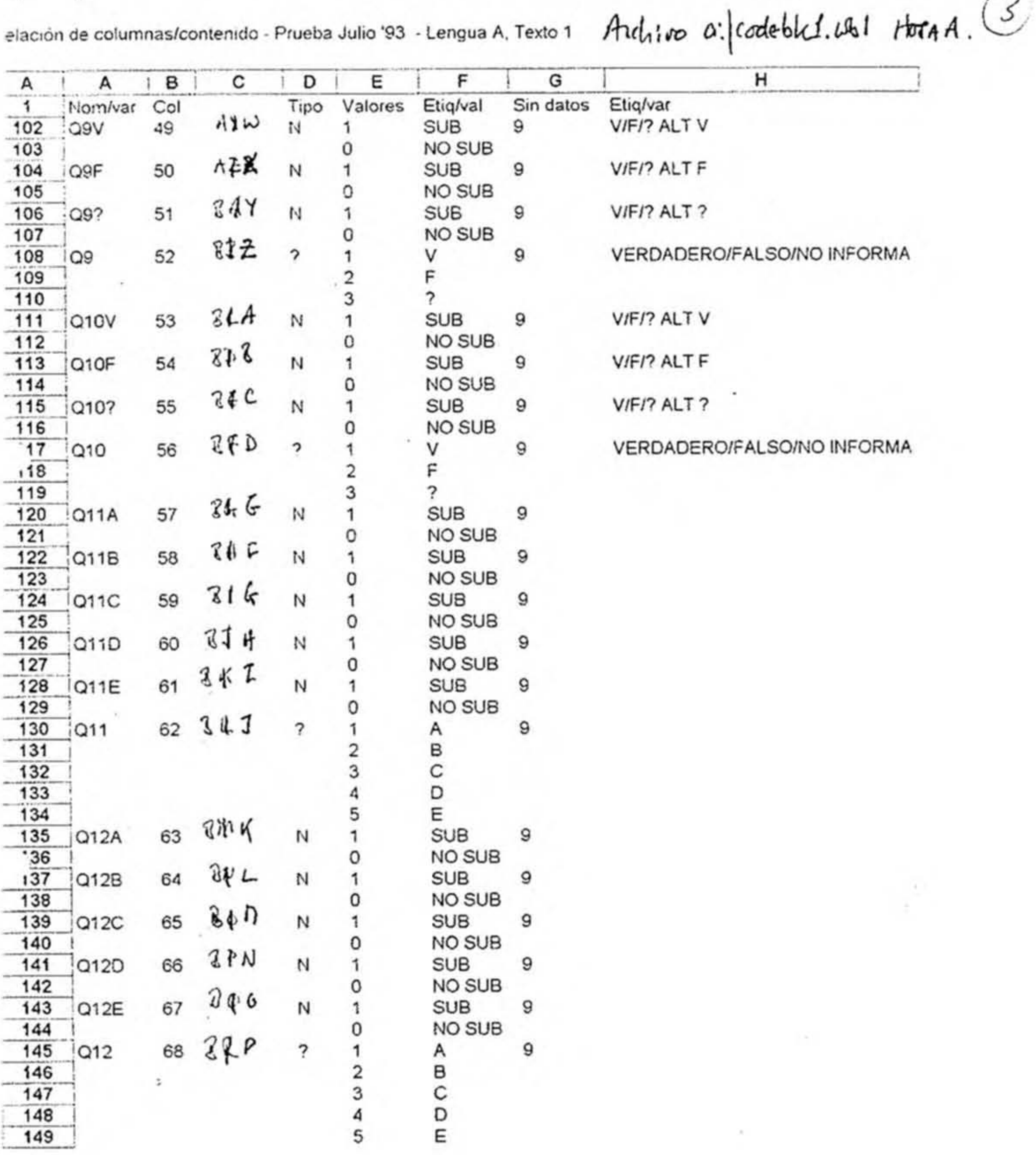
elación de columnas/contenido - Prueba Julio '93 - Lengua A, Texto 1 Archivo a:\ codeldc1.wb1 HOJA.

A	A	B	C	D	E	F.	G	H
1	Nom/var	Col	Celda.	Tipo	Valores	Etiq/val	Sin datos	Etiq/var
3	ID	1-4	B.A5..D5	N	------	-----	9999	Nº SUJETO
4	LE	5-6	B:E5..F5	C	------	-----	99	LENGUA
5	TX	7-8	B:G5 H5	N	------	-----	99	Nº TEXTO
6	Q1A	9	B:I5	N	1	SUB	9	LECTURA RÁPIDA Q1, ALT A
7					0	NO SUB		
8	Q1B	10	B:J5	N	1	SUB	9	LECTURA RÁPIDA Q1, ALT B
9					0	NO SUB		
10	Q1C	11	B:K5	N	1	SUB	9	LECTURA RÁPIDA Q1, ALT C
11					0	NO SUB		
12	Q1D	12	B:L5	N	1	SUB	9	LECTURA RÁPIDA Q1, ALT D
13					0	NO SUB		
14	Q1E	13	B:M5	N	1	SUB	9	LECTURA RÁPIDA Q1, ALT E
15					0	NO SUB		
16	Q1	14	B:N5	?	1	A	9	LECTURA RÁPIDA 1
17					2	B		
.8					3	C		
19					4	D		
20					5	E		
21	Q2A	15	B:K5 O	N	1	SUB	9	LECTURA RÁPIDA Q2, ALT A
22					0	NO SUB		
23	Q2B	16	B:R5 P	N	1	SUB	9	LECTURA RÁPIDA Q2, ALT B
24					0	NO SUB		
25	Q2C	17	B:K5 Q	N	1	SUB	9	LECTURA RÁPIDA Q2, ALT C
26					0	NO SUB		
27	Q2D	18	B:R	N	1	SUB	9	LECTURA RÁPIDA Q2, ALT D
28					0	NO SUB		
29	Q2E	19	B:S5	N	1	SUB	9	LECTURA RÁPIDA Q2, ALT E
30					0	NO SUB		
31	Q2	20	B:T5	?	1	A	9	LECTURA RÁPIDA Q2
32					2	B		
33					3	C		
34					4	D		
35					5	E		
36	Q3A	21	U	N	1	SUB	9	LECTURA RÁPIDA Q3, ALT A
37					0	NO SUB		
38	Q3B	22	V	N	1	SUB	9	LECTURA RÁPIDA Q3, ALT B
39					0	NO SUB		
40	Q3C	23	W	N	1	SUB	9	LECTURA RÁPIDA Q3, ALT C
41					0	NO SUB		
42	Q3D	24	X	N	1	SUB	9	LECTURA RÁPIDA Q3, ALT D
43					0	NO SUB		
44	Q3E	25	Y	N	1	SUB	9	LECTURA RÁPIDA Q3, ALT E
45					0	NO SUB		
46	Q3	26	Z	?	1	A	9	LECTURA RÁPIDA Q3
47					2	B		
48					3	C		
49					4	D		
50					5	E		
~~51~~	~~Q4A~~	~~27~~	~~AA~~	~~N~~	~~1~~	~~SUB~~	~~9~~	~~LECTURA RÁPIDA Q4, ALT A~~
~~52~~					~~0~~	~~NO SUB~~		
~~53~~	~~Q4B~~	~~28~~	~~AB~~	~~N~~	~~1~~	~~SUB~~	~~9~~	~~LECTURA RÁPIDA Q4, ALT B~~

(2.

Relación de columnas/contenido - Prueba Julio '93 - Lengua A, Texto 1 Archivo a:\codebk1.wb1 Hoja A.

A	A	B	C	D	E	F	G	H
1	Nom/var	Col		Tipo	Valores	Etiq/val	Sin datos	Etiq/var
51	Q4A	27	AA	N	1	SUB	9	LECTURA RÁPIDA Q4, ALT A
52					0	NO SUB		
53	Q4B	28	AB	N	1	SUB	9	LECTURA RÁPIDA Q4, ALT B
54					0	NO SUB		
55	Q4C	29	AC	N	1	SUB	9	LECTURA RÁPIDA Q4, ALT C
56					0	NO SUB		
57	Q4D	30	AD	N	1	SUB	9	LECTURA RÁPIDA Q4, ALT D
58					0	NO SUB		
59	Q4E	31	AE	N	1	SUB	9	LECTURA RÁPIDA Q4, ALT E
60					0	NO SUB		
61	Q4	32	AF	?	1	A	9	LECTURA RÁPIDA Q4
62					2	B		
63					3	C		
64					4	D		
65					5	E		
66	Q5V	33	AG	N	1	SUB	9	V/F/? ALT V
67					0	NO SUB		
68	Q5F	34	AH	N	1	SUB	9	V/F/? ALT F
69					0	NO SUB		
70	Q5?	35	AI	N	1	SUB	9	V/F/? ALT ?
71					0	NO SUB		
72	Q5	36	AJ	?	1	V	9	VERDADERO/FALSO/NO INFORMA
73					2	F		
74					3	?		
75	Q6V	37	AK	N	1	SUB	9	V/F/? ALT V
76					0	NO SUB		
77	Q6F	38	AL	N	1	SUB	9	V/F/? ALT F
78					0	NO SUB		
79	Q6?	39	AM	N	1	SUB	9	V/F/? ALT ?
80					0	NO SUB		
81	Q6	40	AN	?	1	V	9	VERDADERO/FALSO/NO INFORMA
82					2	F		
83					3	?		
84	Q7V	41	AO	N	1	SUB	9	V/F/? ALT V
85					0	NO SUB		
86	Q7F	42	AP	N	1	SUB	9	V/F/? ALT F
87					0	NO SUB		
88	Q7?	43	AQ	N	1	SUB	9	V/F/? ALT ?
89					0	NO SUB		
90	Q7	44	AR	?	1	V	9	VERDADERO/FALSO/NO INFORMA
91					2	F		
92					3	?		
93	Q8V	45	AS	N	1	SUB	9	V/F/? ALT V
94					0	NO SUB		
95	Q8F	46	AT	N	1	SUB	9	V/F/? ALT F
96					0	NO SUB		
97	Q8?	47	AU	N	1	SUB	9	V/F/? ALT ?
98					0	NO SUB		
99	Q8	48	AV	?	1	V	9	VERDADERO/FALSO/NO INFORMA
100					2	F		
101					3	?		



elación de columnas/contenido - Prueba Julio '93 - Lengua A, Texto 1 Archivo a:\codebk1.wb1 Hoja A. (3)

A	A	B	C	D	E	F	G	H
1	Nom/var	Col		Tipo	Valores	Etiq/val	Sin datos	Etiq/var
102	Q9V	49	AYW	N	1	SUB	9	V/F/? ALT V
103					0	NO SUB		
104	Q9F	50	AZX	N	1	SUB	9	V/F/? ALT F
105					0	NO SUB		
106	Q9?	51	BAY	N	1	SUB	9	V/F/? ALT ?
107					0	NO SUB		
108	Q9	52	BBZ	?	1	V	9	VERDADERO/FALSO/NO INFORMA
109					2	F		
110					3	?		
111	Q10V	53	BCA	N	1	SUB	9	V/F/? ALT V
112					0	NO SUB		
113	Q10F	54	BDB	N	1	SUB	9	V/F/? ALT F
114					0	NO SUB		
115	Q10?	55	BEC	N	1	SUB	9	V/F/? ALT ?
116					0	NO SUB		
17	Q10	56	BFD	?	1	V	9	VERDADERO/FALSO/NO INFORMA
18					2	F		
119					3	?		
120	Q11A	57	BGE	N	1	SUB	9	
121					0	NO SUB		
122	Q11B	58	BHF	N	1	SUB	9	
123					0	NO SUB		
124	Q11C	59	BIG	N	1	SUB	9	
125					0	NO SUB		
126	Q11D	60	BJH	N	1	SUB	9	
127					0	NO SUB		
128	Q11E	61	BKI	N	1	SUB	9	
129					0	NO SUB		
130	Q11	62	BLJ	?	1	A	9	
131					2	B		
132					3	C		
133					4	D		
134					5	E		
135	Q12A	63	BMK	N	1	SUB	9	
36					0	NO SUB		
137	Q12B	64	BNL	N	1	SUB	9	
138					0	NO SUB		
139	Q12C	65	BOM	N	1	SUB	9	
140					0	NO SUB		
141	Q12D	66	BPN	N	1	SUB	9	
142					0	NO SUB		
143	Q12E	67	BQO	N	1	SUB	9	
144					0	NO SUB		
145	Q12	68	BRP	?	1	A	9	
146					2	B		
147					3	C		
148					4	D		
149					5	E		

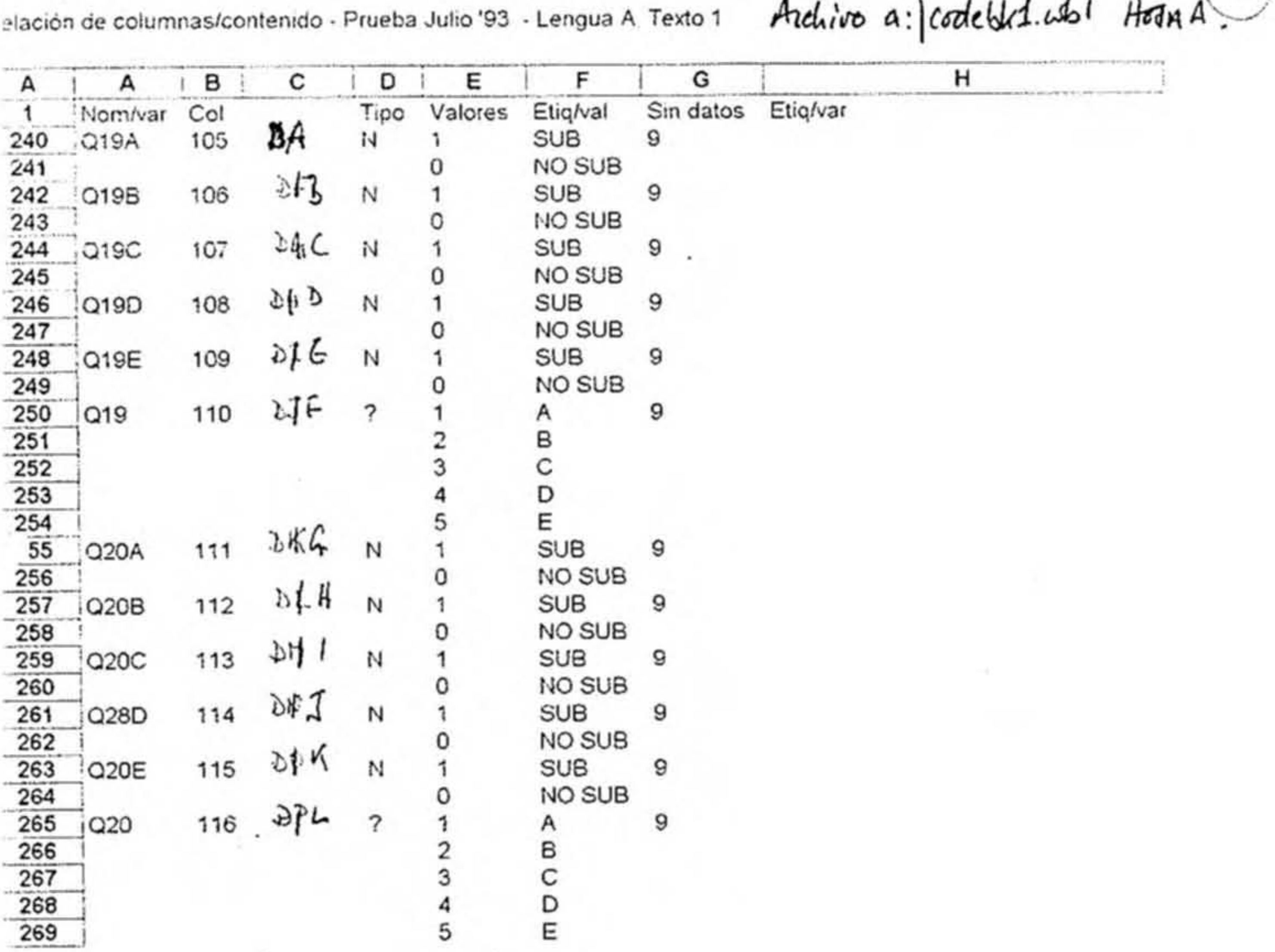
elación de columnas/contenido - Prueba Julio '93 - Lengua A, Texto 1 Archivo a:\codebk1.wb1 Hoja A. (4

A	A	B	C	D	E	F	G	H
1	Nom/var	Col		Tipo	Valores	Etiq/val	Sin datos	Etiq/var
150	Q13A	69	BQ	N	1	SUB	9	
151					0	NO SUB		
152	Q13B	70	BR	N	1	SUB	9	
153					0	NO SUB		
154	Q13C	71	BS	N	1	SUB	9	
155					0	NO SUB		
156	Q13D	72	BT	N	1	SUB	9	
157					0	NO SUB		
158	Q13E	73	BU	N	1	SUB	9	
159					0	NO SUB		
160	Q13	74	BV	?	1	A	9	
161					2	B		
162					3	C		
163					4	D		
164					5	E		
165	Q14A	75	BW	N	1	SUB	9	
66					0	NO SUB		
167	Q14B	76	BX	N	1	SUB	9	
168					0	NO SUB		
169	Q14C	77	CY	N	1	SUB	9	
170					0	NO SUB		
171	Q14D	78	CZ	N	1	SUB	9	
172					0	NO SUB		
173	Q14E	79	CA	N	1	SUB	9	
174					0	NO SUB		
175	Q14	80	CB	?	1	A	9	
176					2	B		
177					3	C		
178					4	D		
179					5	E		
180	Q15A	81	CC	N	1	SUB	9	
181					0	NO SUB		
182	Q15B	82	CD	N	1	SUB	9	
183					0	NO SUB		
184	Q15C	83	CG	N	1	SUB	9	
85					0	NO SUB		
186	Q15D	84	CF	N	1	SUB	9	
187					0	NO SUB		
188	Q15E	85	CG	N	1	SUB	9	
189					0	NO SUB		
190	Q15	86	CH	?	1	A	9	
191					2	B		
192					3	C		
193					4	D		
194					5	E		

5

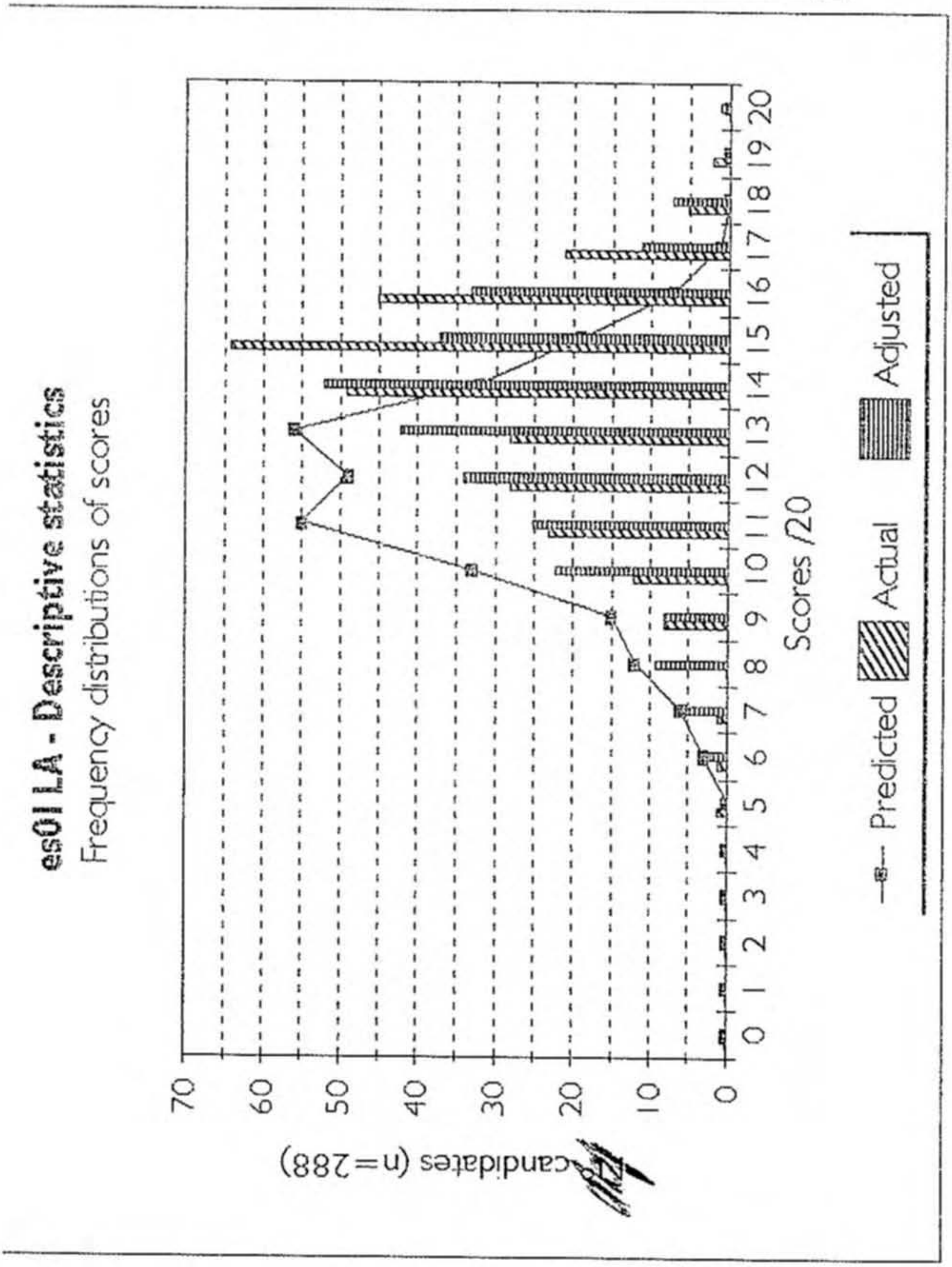
lacion de columnas/contenido - Prueba Julio '93 - Lengua A, Texto 1 Archivo a:\codebk1.wb1 Hoja A

A	A	B	C	D	E	F	G	H
1	Nom/var	Col		Tipo	Valores	Etiq/val	Sin datos	Etiq/var
195	Q16A	87	CLI	N	1	SUB	9	
196					0	NO SUB		
197	Q16B	88	CMJ	N	1	SUB	9	
198					0	NO SUB		
199	Q16C	89	CNK	N	1	SUB	9	
200					0	NO SUB		
201	Q16D	90	COL	N	1	SUB	9	
202					0	NO SUB		
203	Q16E	91	CPM	N	1	SUB	9	
204					0	NO SUB		
205	Q16	92	CQN	?	1	A	9	
206					2	B		
207					3	C		
208					4	D		
209					5	E		
210	Q17A	93	CRO	N	1	SUB	9	
211					0	NO SUB		
212	Q17B	94	CSP	N	1	SUB	9	
213					0	NO SUB		
214	Q17C	95	CTQ	N	1	SUB	9	
215					0	NO SUB		
216	Q17D	96	CUR	N	1	SUB	9	
217					0	NO SUB		
218	Q17E	97	CVS	N	1	SUB	9	
219					0	NO SUB		
220	Q17	98	CWT	?	1	A	9	
221					2	B		
222					3	C		
223					4	D		
224					5	E		
225	Q18A	99	CXU	N	1	SUB	9	
226					0	NO SUB		
227	Q18B	100	CYV	N	1	SUB	9	
228					0	NO SUB		
229	Q18C	101	CZW	N	1	SUB	9	
230					0	NO SUB		
231	Q18D	102	DAX	N	1	SUB	9	
232					0	NO SUB		
233	Q18E	103	DBY	N	1	SUB	9	
234					0	NO SUB		
235	Q18	104	DCZ	?	1	A	9	
236					2	B		
237					3	C		
238					4	D		
239					5	E		



elación de columnas/contenido - Prueba Julio '93 - Lengua A Texto 1 Archivo a:\codeblk1.wb1 Hoja A 6

A	A	B	C	D	E	F	G	H
1	Nom/var	Col		Tipo	Valores	Etiq/val	Sin datos	Etiq/var
240	Q19A	105	BA	N	1	SUB	9	
241					0	NO SUB		
242	Q19B	106	D|B	N	1	SUB	9	
243					0	NO SUB		
244	Q19C	107	D|C	N	1	SUB	9	
245					0	NO SUB		
246	Q19D	108	D|D	N	1	SUB	9	
247					0	NO SUB		
248	Q19E	109	D|E	N	1	SUB	9	
249					0	NO SUB		
250	Q19	110	D|F	?	1	A	9	
251					2	B		
252					3	C		
253					4	D		
254					5	E		
55	Q20A	111	D|G	N	1	SUB	9	
256					0	NO SUB		
257	Q20B	112	D|H	N	1	SUB	9	
258					0	NO SUB		
259	Q20C	113	D|I	N	1	SUB	9	
260					0	NO SUB		
261	Q28D	114	D|J	N	1	SUB	9	
262					0	NO SUB		
263	Q20E	115	D|K	N	1	SUB	9	
264					0	NO SUB		
265	Q20	116	D|L	?	1	A	9	
266					2	B		
267					3	C		
268					4	D		
269					5	E		


es01 LA - Descriptive statistics
Frequency distributions of scores
candidates (n=288)
0
10
20
30
40
50
60
70
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Scores /20
Predicted
Actual
Adjusted

PREDICTED

En la prueba es01 las alternativas previstas fueron las siguientes:

1.	**A**	B	C	D	E
2.	A	B	**C**	D	E
3.	A	**B**	C	D	E
4.	A	B	C	**D**	E
5.	V	F	**?**		
6.	V	**F**	?		
7.	**V**	F	?		
8.	V	F	**?**		
9.	V	**F**	?		
10.	V	F	**?**		
11.	**A**	B	C	D	E
12.	**A**	B	C	D	E
13.	A	B	**C**	D	E
14.	A	**B**	C	D	E
15.	A	**B**	C	D	E
16.	A	B	C	D	**E**
17.	A	B	**C**	D	E
18.	A	B	C	**D**	E
19.	A	**B**	C	D	E
20.	A	B	C	**D**	E

ACTUAL

En la prueba es01 las respuestas actuales con relación al índice de dificultad fueron:

1.	**A**	B	C	D	E
2.	A	B	**C**	D	E
3.	A	**B**	C	D	E
4.	A	B	C	**D**	E
5.	**V**	F	?		
6.	V	**F**	**?**		
7.	V	**F**	?		
8.	**V**	F	?		
9.	V	**F**	?		
10.	**V**	**F**	**?**		
11.	A	**B**	C	D	E
12.	**A**	B	C	D	E
13.	A	B	C	D	**E**
14.	A	**B**	C	D	E
15.	A	**B**	C	D	E
16.	A	B	C	D	**E**
17.	A	B	**C**	D	E
18.	A	B	C	**D**	E
19.	A	**B**	C	D	E
20.	A	B	C	**D**	E

ADJUSTED

En la prueba es01 las respuestas revisadas

1.	**A**	B	C	D	E
2.	A	B	**C**	D	E
3.	A	**B**	C	D	E
4.	A	B	C	**D**	E
5.	V	F	**?**		
6.	V	**F**	?		
7.	**V**	F	?		
8.	**V**	F	?		
9.	V	**F**	?		
10.	V	F	**?**		
11.	A	**B**	C	D	E
12.	**A**	B	C	D	E
13.	A	B	C	D	**E**
14.	A	**B**	C	D	E
15.	A	**B**	C	D	E
16.	A	B	C	D	**E**
17.	A	B	**C**	D	E
18.	A	B	C	**D**	E
19.	A	**B**	C	D	E
20.	A	B	C	**D**	E

es01 Facility values (FV)

Candidates (n =	288								
Item	1A	1B	1C	1D	1E	*p*	*q*	*pq*	*Sum pq*
Actual responses	282	1	2	2	0	0.98	0.02	0.02	3.54
FV	0.98	0.00	0.01	0.01	0.00				
Total responses	287								
Errors	1								
Errors/Total (%)	0.3								

Item	2A	2B	2C	2D	2E	*p*	*q*	*pq*
Actual responses	2	1	280	0	3	0.97	0.03	0.03
FV	0.01	0.00	0.97	0.00	0.01			
Total responses	286							
Errors	2							
Errors/Total (%)	0.7							

Item	3A	3B	3C	3D	3E	*p*	*q*	*pq*
Actual responses	0	286	0	0	0	0.99	0.01	0.01
FV	0.00	0.99	0.00	0.00	0.00			
Total responses	286							
Errors	2							
Errors/Total (%)	0.7							

Item	4A	4B	4C	4D	4E	*p*	*q*	*pq*
Actual responses	0	0	0	245	45	0.85	0.15	0.13
FV	0.00	0.00	0.00	0.85	0.16			
Total responses	290							
Errors	(2)							
Errors/Total (%)	(0.7)							

Item	5T	5F	5?			*p*	*q*	*pq*
Actual responses	184	43	62			0.22	0.78	0.17
FV	0.64	0.15	0.22					
Total responses	289							
Errors	(1)							
Errors/Total (%)	(0.3)							

es01 Facility values (FV)

Item	6T	6F	6?	p	q	pq
Actual responses	30	135	122	0.47	0.53	0.25
FV	0.10	0.47	0.42			
Total responses	287					
Errors	1					
Errors/Total (%)	0.3					

Item	7T	7F	7?	p	q	pq
Actual responses	99	149	39	0.34	0.66	0.23
FV	0.34	0.52	0.14			
Total responses	287					
Errors	1					
Errors/Total (%)	0.3					

Item	8T	8F	8?	p	q	pq
Actual responses	153	105	30	0.36	0.64	0.23
FV	0.53	0.36	0.10			
Total responses	288					
Errors	0					
Errors/Total (%)	0.0					

Item	9T	9F	9?	p	q	pq
Actual responses	40	215	33	0.75	0.25	0.19
FV	0.14	0.75	0.11			
Total responses	288					
Errors	0					
Errors/Total (%)	0.0					

Item	10T	10F	10?	p	q	pq
Actual responses	85	106	96	0.33	0.67	0.22
FV	0.30	0.37	0.33			
Total responses	287					
Errors	1					
Errors/Total (%)	0.3					

es01 Facility values (FV)

Item	I1A	I1B	I1C	I1D	I1E	*p*	*q*	*pq*
Actual responses	37	179	44	4	25	0.62	0.38	0.24
FV	0.13	0.62	0.15	0.01	0.09			
Total responses	289							
Errors	(1)							
Errors/Total (%)	(0.3)							

Item	I2A	I2B	I2C	I2D	I2E	*p*	*q*	*pq*
Actual responses	239	11	6	9	25	0.83	0.17	0.14
FV	0.83	0.04	0.02	0.03	0.09			
Total responses	290							
Errors	(2)							
Errors/Total (%)	(0.7)							

Item	I3A	I3B	I3C	I3D	I3E	*p*	*q*	*pq*
Actual responses	80	1	15	14	178	0.62	0.38	0.24
FV	[redacted]	0.00	0.05	0.05	0.62			
Total responses	288							
Errors	0							
Errors/Total (%)	0.0							

Item	I4A	I4B	I4C	I4D	I4E	*p*	*q*	*pq*
Actual responses	54	193	17	14	8	0.67	0.33	0.22
FV	[redacted]	0.67	0.06	0.05	0.03			
Total responses	286							
Errors	2							
Errors/Total (%)	0.7							

Item	I5A	I5B	I5C	I5D	I5E	*p*	*q*	*pq*
Actual responses	4	240	35	1	7	0.83	0.17	0.14
FV	0.01	0.83	0.12	0.00	0.02			
Total responses	287							
Errors	1							
Errors/Total (%)	0.3							

es01 Facility values (FV)

Item	16A	16B	16C	16D	16E	*p*	*q*	*pq*
Actual responses	20	14	19	4	231	0.80	0.20	0.16
FV	0.07	0.05	0.07	0.01	0.80			
Total responses	288							
Errors	0							
Errors/Total (%)	0.0							

Item	17A	17B	17C	17D	17E	*p*	*q*	*pq*
Actual responses	3	8	176	2	95	0.61	0.39	0.24
FV	0.01	0.03	0.61	0.01	[redacted]			
Total responses	284							
Errors	4							
Errors/Total (%)	1.4							

Item	18A	18B	18C	18D	18E	*p*	*q*	*pq*
Actual responses	34	21	20	190	21	0.66	0.34	0.22
FV	0.12	0.07	0.07	0.66	0.07			
Total responses	286							
Errors	2							
Errors/Total (%)	0.7							

Item	19A	19B	19C	19D	19E	*p*	*q*	*pq*
Actual responses	10	173	3	42	54	0.60	0.40	0.24
FV	0.03	0.60	0.01	[redacted]	[redacted]			
Total responses	282							
Errors	6							
Errors/Total (%)	2.1							

Item	20A	20B	20C	20D	20E	*p*	*q*	*pq*
Actual responses	14	38	25	180	27	0.63	0.38	0.23
FV	0.05	0.13	0.09	0.63	0.09			
Total responses	284							
Errors	4							
Errors/Total (%)	1.4							

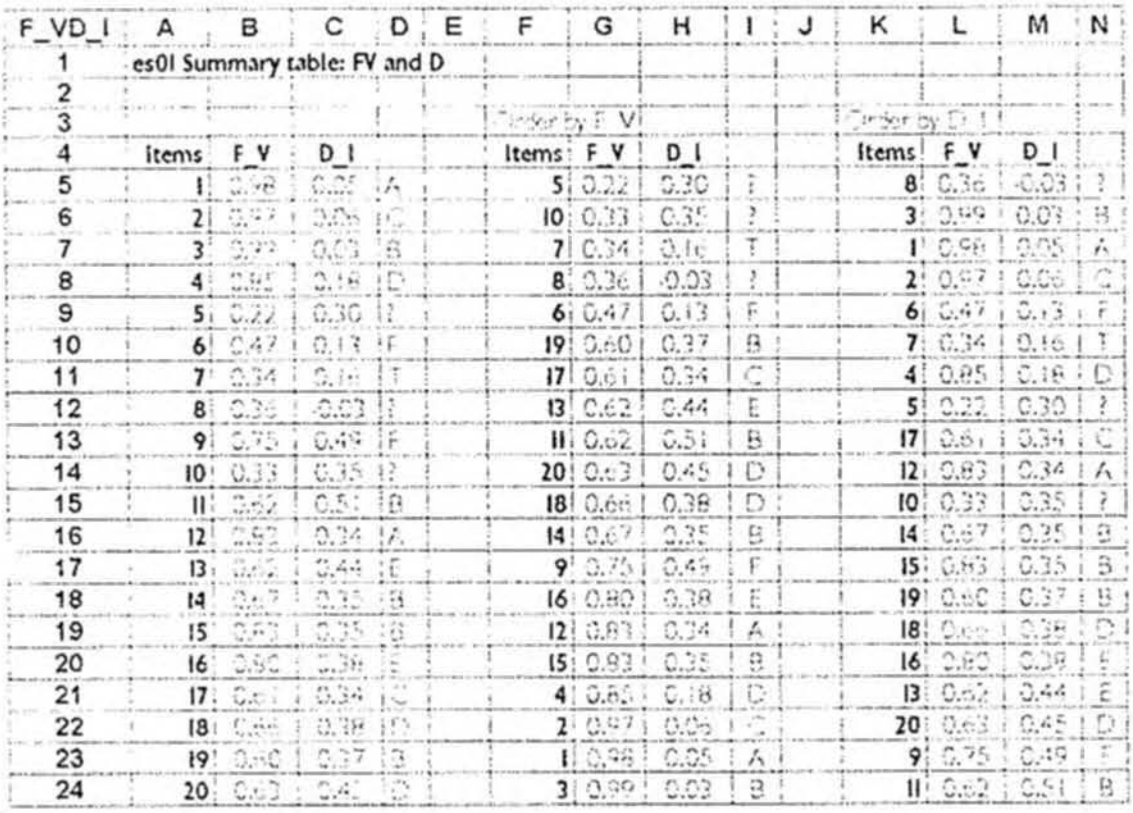
~~29/12/95~~

F K .L N

	es01 índice de discriminación																				
	población =	288																			
4		Q1	Q2	Q3	Q4	Q5V	Q6V	Q7V	Q8V	Q9V	Q10V	Q11A	Q12	Q13A	Q14	Q15	Q16	Q17	Q18	Q19	Q20
	Correcto 'mejores'	78	78	79	75	39	5	35	42	0	17	4	75	11	66	77	75	60	66	64	66
	Correcto 'peores'	74	73	77	61	63	16	22	39	27	34	15	48	36	38	49	45	33	36	35	30
	Nº 0 27,5%	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79	79
	Discriminación	0,05	0,06	0,03	0,18	-0,30	-0,14	0,16	0,04	-0,34	-0,21	-0,14	0,34	-0,32	0,35	0,35	0,38	0,34	0,38	0,37	0,45
						Q5F	Q6F	Q7F	Q8F	Q9F	Q10F	Q11B		Q13B							
	Correcto 'mejores'					8	42	28	30	77	19	66		0							
	Correcto 'peores'					8	32	47	32	38	31	26		0							
	Nº 0 27,5%					79	79	79	79	79	79	79		79							
	Discriminación					0,00	0,13	-0,24	-0,03	0,49	-0,15	0,51		0,00							
						Q5?	Q6?	Q7?	Q8?	Q9?	Q10?	Q11C		Q13C							
	Correcto 'mejores'					32	32	16	7	2	42	3		4							
	Correcto 'peores'					8	30	9	8	14	14	29		7							
	Nº 0 27,5%					79	79	79	79	79	79	79		79							
20	Discriminación					0,30	0,03	0,09	-0,01	-0,15	0,35	-0,33		-0,04							
												Q11D		Q13D							
	Correcto 'mejores'											1		2							
	Correcto 'peores'											3		8							
	Nº 0 27,5%											79		79							
	Discriminación											-0,03		-0,08							
												Q11E		Q13E							
	Correcto 'mejores'											5		62							
	Correcto 'peores'											6		27							
	Nº 0 27,5%											79		79							
32	Discriminación											-0,01		0,44							

I 21/8/96.

c:\document\thesis\results\stats\es01\es01stat.wb1 Hoja ~~1~~ Índice de discriminación

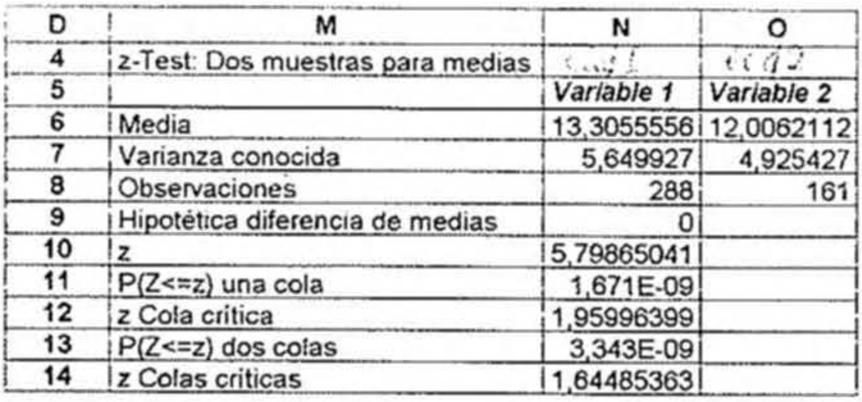


F_VD_I	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	es0l Summary table: FV and D													
2														
3						Order by F V					Order by D I			
4	Items	F_V	D_I			Items	F_V	D_I			Items	F_V	D_I	
5	1	0.98	0.05	A		5	0.22	0.30	?		8	0.36	-0.03	?
6	2	0.97	0.06	C		10	0.33	0.35	?		3	0.99	0.03	B
7	3	0.99	0.03	B		7	0.34	0.16	T		1	0.98	0.05	A
8	4	0.85	0.18	D		8	0.36	-0.03	?		2	0.97	0.06	C
9	5	0.22	0.30	?		6	0.47	0.13	F		6	0.47	0.13	F
10	6	0.47	0.13	F		19	0.60	0.37	B		7	0.34	0.16	T
11	7	0.34	0.16	T		17	0.61	0.34	C		4	0.85	0.18	D
12	8	0.36	-0.03	?		13	0.62	0.44	E		5	0.22	0.30	?
13	9	0.75	0.49	F		11	0.62	0.51	B		17	0.61	0.34	C
14	10	0.33	0.35	?		20	0.63	0.45	D		12	0.83	0.34	A
15	11	0.62	0.51	B		18	0.66	0.38	D		10	0.33	0.35	?
16	12	0.83	0.34	A		14	0.67	0.35	B		14	0.67	0.35	B
17	13	0.62	0.44	E		9	0.75	0.49	F		15	0.83	0.35	B
18	14	0.67	0.35	B		16	0.80	0.38	E		19	0.60	0.37	B
19	15	0.83	0.35	B		12	0.83	0.34	A		18	0.66	0.38	D
20	16	0.80	0.38	E		15	0.83	0.35	B		16	0.80	0.38	E
21	17	0.61	0.34	C		4	0.85	0.18	D		13	0.62	0.44	E
22	18	0.66	0.38	D		2	0.97	0.06	C		20	0.63	0.45	D
23	19	0.60	0.37	B		1	0.98	0.05	A		9	0.75	0.49	F
24	20	0.63	0.45	D		3	0.99	0.03	B		11	0.62	0.51	B

21/8/96

es01 Summary table		
Item	FV	D
Q1	0.98	0.05
Q2	0.97	0.06
Q3	0.99	0.03
Q4	0.85	0.18
Q5	0.22	0.30
Q6	0.47	-0.13
Q7	0.34	-0.16
Q8	0.53	-0.04
Q9	0.75	0.49
Q10	0.33	-0.35
Q11	0.62	0.51
Q12	0.83	0.34
Q13	0.62	0.44
Q14	0.67	0.35
Q15	0.83	0.35
Q16	0.80	0.38
Q17	0.61	0.34
Q18	0.66	0.38
Q19	0.60	0.37
Q20	0.63	0.45

z-Test es01 & es02



D	M	N	O
4	z-Test: Dos muestras para medias	es01	es02
5		*Variable 1*	*Variable 2*
6	Media	13,3055556	12,0062112
7	Varianza conocida	5,649927	4,925427
8	Observaciones	288	161
9	Hipotética diferencia de medias	0	
10	z	5,79865041	
11	P(Z<=z) una cola	1,671E-09	
12	z Cola crítica	1,95996399	
13	P(Z<=z) dos colas	3,343E-09	
14	z Colas críticas	1,64485363	

c:\document\thesis\results\stats\es01\es01stat.wb1 Hoja [illegible]

Sub-test es02

8 / EDUCACIÓN

LA ADQUISICIÓN DE VALORES

Urge la tolerancia

Unicef intensifica su campaña para educar a los jóvenes como 'ciudadanos del mundo'

ANA FERNÁNDEZ

A El Fondo de las Naciones Unidas para la Infancia (Unicef) está llevando a cabo una campaña internacional dirigida a concienciar e integrar en los sistemas educativos una nueva opción pedagógica que es la educación para el desarrollo, y cuyo objetivo fundamental es sensibilizar al ser humano a fin de que colabore en la solución de los problemas y diferencias, tanto económicas como culturales o étnicas, que afectan a nuestra sociédad, ofreciendo así una alternativa frente a la violencia que a menudo se utiliza para dirimir tales diferencias.

B El responsable del Programa de Educación para el Desarrollo de la oficina europea de la Unicef, Andrés Guerrero Feliú, opina que, aunque la educación para el desarrollo no da una respuesta a todos los problemas que surgen en este mundo interrelacionado, "intenta, al menos, crear en los jóvenes las actitudes que les van a permitir comprender los complejos sistemas de relaciones que dominan nuestra sociedad y de cómo éstos influyen en uno mismo en su relación con el entorno".

C La educación para el desarrollo, una de las principales preocupaciones del fondo desde los años 70, está propulsada principalmente, desde los propios comités nacionales que la Unicef tiene en la casi totalidad de los paises industrializados que son los que más elementos de juicio tienen a la hora de elaborar material informativo y didáctico adaptado a cada país.

D Si en un principio esta inquietud nació para informar y sensibilizar a los jóvenes del Norte sobre los problemas de los paises en desarrollo, el nuevo enfoque de la organización está centrado en la promoción de una "ciudadanía global".

E La solidaridad, la paz, la tolerancia, la justicia social, la conciencia de los problemas del medio ambiente son los valores que promueve la educación para el desarrollo. Asimismo, intenta ofrecer a los jóvenes los conocimientos necesarios que les permitan poner en práctica dichos valores, según Guerrero.

F "La guerra es un producto de la mente del ser humano. Si somos capaces de generar violencia, también lo somos de crear armonía. Frente a nuestra capacidad de destrucción debemos explotar la capacidad de construcción".

G Opina que la crisis en la antigua Yugoslavia refleja "el fracaso de la no práctica de la educación para el desarrollo". Esto nos enseña que los problemas "se resuelven sobre todo a nivel local en su interacción con el medio donde se producen".

Ciudadanía global

H Los tejidos de relaciones humanas, que en el mundo de hoy se multiplican aceleradamente gracias al avance de los sistemas de comunicación, hacen que el planeta se enfrente a un complejo proceso de interconexiones que generan situaciones y problemas de naturaleza global.

I Por ello, la Unicef considera imprescindible educar a la juventud y fomentar una visión global, mediante el estudio del planeta y de su gente, de los distintos estilos de vida y de la forma en que se aprehende la realidad mundial.

J En este sentido, los conceptos de interdependencia —aprender a valorar las interconexiones y las consecuencias que muchas acciones efectuadas a nivel local acarrean a nivel mundial—, la exploración de otros modos de vida y puntos de vista, eliminando los estereotipos que son siempre juicios de valor muy simplistas o la justicia social, mediante la adquisición de conocimientos en materia de derechos humanos y su aplicación a la vida diaria, son elementos que forman parte del aprendizaje de la ciudadanía global.

K Asimismo, Guerrero destaca otros conceptos inherentes a este nuevo tipo de ser humano como son la enseñanza a la juventud del origen y la solución de conflictos, para lo que necesita comprender las diferentes fuentes y las causas de los mismos. Debe aprender a luchar por la paz y entender que las medidas que se toman hoy afectarán a las generaciones futuras.

L "Una de las dimensiones de la globalidad es el tiempo", señala; cada acontecimiento hunde sus raíces en algo que ya se produjo y a su vez influirá en lo que se va a producir. Una preparación para el futuro implica comprender que cada uno de nosotros tiene algo que decir ante las fuerzas que provocan los cambios y una participación activa en el mismo proceso".

INFORMACIÓN SOBRE EL EXAMEN

Este examen consta de dos partes que son pruebas de habilidades de lectura.

Antes de comenzar comprueba haber rellenado los datos personales en las hojas de respuestas, y haber entendido bien como señalar las respuestas correctas.

La primera prueba es un texto en lengua española; la segunda es un texto en alemán, francés, o inglés.

El tiempo total de duración del examen es de 2 horas. La distribución del tiempo se deja al arbitrio de cada uno, pero se sugiere emplear alrededor de una hora para cada parte.

Está prohibido salirse del examen sin entregar los papeles. En cualquier caso no se permitirá la salida durante los últimos diez minutos del examen.

No se admitirá el uso de ningún libro, papel o cualquier otro tipo de ayuda. Las personas que los utilicen se descalificarán automáticamente.

Se recomienda el uso de lápiz y goma de borrar para el examen. No es necesario el uso de bolígrafo.

Se puede usar la hoja de preguntas como borrador, pero todas las respuestas deberán consignarse en las hojas de respuestas.

El formato de las preguntas es de elección múltiple. Las instrucciones son de la siguiente manera:

> Lee el texto y señala a cuál de los párrafos podría corresponder cada una de las siguientes frases como título. Subraya la letra en la hoja de respuestas.
>
> y
>
> V (Verdadero), F (Falso) o ? (el texto no informa sobre este punto), y subraya la letra o el símbolo correspondiente en la hoja de respuestas.
>
> y
>
> Elige la opción más apropiada y subraya la letra en la hoja de respuestas.

Si se comete algún error al rellenar la hoja de respuestas borra la respuesta errónea y subraya la correcta, o tacha con una "X" la respuesta errónea y subraya la correcta.

Al final del examen se entregarán todos los papeles: es decir, textos, hojas de preguntas, y hojas de respuestas.

Español - Prueba de habilidades de lectura - Texto Nº 2

1er APELLIDO:

2º APELLIDO:

NOMBRE:

DNI:

[SCORE: □□]

Subraya la letra o símbolo que corresponda.

1.	A	B	C	D	E	F	G	H	I	J	K	L
2.	A	B	C	D	E	F	G	H	I	J	K	L
3.	A	B	C	D	E	F	G	H	I	J	K	L
4.	A	B	C	D	E	F	G	H	I	J	K	L
5.	A	B	C	D	E	F	G	H	I	J	K	L
6.	V	F	?									
7.	V	F	?									
8.	V	F	?									
9.	V	F	?									
10.	V	F	?									
11.	A	B	C	D	E							
12.	A	B	C	D	E							
13.	A	B	C	D	E							
14.	A	B	C	D	E							
15.	A	B	C	D	E							
16.	A	B	C	D	E							
17.	A	B	C	D	E							
18.	A	B	C	D	E							
19.	A	B	C	D	E							
20.	A	B	C	D	E							

Español - Prueba de habilidades de lectura - Texto Nº 2

"Urge la tolerancia"

Lee el texto y señala a cuál de los párrafos podría corresponder cada una de las siguientes frases como título. Subraya la letra en la hoja de respuestas.

1. "El objetivo fundamental del programa."

2. "Los valores que promueve el programa."

3. "La globalidad de los problemas."

4. "Los derechos humanos."

5. "Las generaciones futuras."

Elige V (Verdadero), F (Falso) o ? (el texto no informa sobre este punto), y subraya la letra o el símbolo correspondiente en la hoja de respuestas.

6. La campaña intenta solucionar los problemas de las personas que utilizan la violencia para dirimir sus diferencias.

7. Guerrero está al frente de la oficina europea de la Unicef.

8. El fondo fue creado en la década de los años setenta.

9. La crisis de la antigua Yugoslavia se deriva en parte de la falta de puesta en práctica de los valores que promueve el fondo.

10. Los gobiernos de los países industrializados invertirán en la elaboración del material didáctico e informativo.

Elige la alternativa más apropiada y subraya la letra en la hoja de respuestas.

11. La educación para el desarrollo persigue la solución de los problemas y diferencias __________ sociedad.

A ... económicas de la ...
B ... económicos que afectan a la ...
C ... económicos y culturales, que afectan a la ...
D ... económicos, culturales y étnicos que afectan a la ...
E ... económicas, culturales y étnicas, que afectan a nuestra ...

12. Guerrero opina que ...

A se pueden comprender las relaciones de dominio de la sociedad.
B los problemas no se solucionan en su interacción con el medio.
C el tiempo es un concepto de dimensiones globales.
D podemos generar algo de más valor que la violencia.
E las fuerzas provocan una participación en el proceso.

13. La frase "intenta, al menos ... con el entorno." (pár. B), sirve como ...

A ilustración.
B matización.
C aviso.
D sugerencia.
E consejo.

14. La educación para el desarrollo es un proyecto ...

A de países industrializados.
B de nueva opción pedagógica.
C de las Naciones Unidas.
D centrado en la infancia.
E en los países del Norte.

15. En la frase "esto nos enseña que los problemas se resuelven sobre todo ..." (pár. G), "esto" ser refiere a la ...

A crisis actual en el antiguo Yugoslavia.

B falta de éxito de la no práctica de la educación para el desarrollo.

C educación para el desarrollo.

D falta de éxito de la educación para el desarrollo.

E falta de éxito de la práctica de la educación para el desarrollo.

16. ¿Qué es mayor?

A Los tejidos de relaciones humanas.
B Los sistemas de comunicación.
C El estudio de los estilos de vida.
D El proceso de interconexiones.
E La explotación de otras vías.

17. La Unicef considera imprescindible fomentar ...

A el estudio de los distintos modos de vida.

B una visión global de las formas de aprehender la realidad mundial.

C la educación de la juventud para aprender de la realidad mundial.

D una visión global del planeta y su gente.

E la educación de la visión de la juventud.

18. Los conceptos inherentes al nuevo tipo de ser humano son los conceptos __________ ser humano.

A ... coaligados al ...
B ... colaterales del ...
C ... circunstanciales del ...
D ... adicionales al ...
E ... esenciales del ...

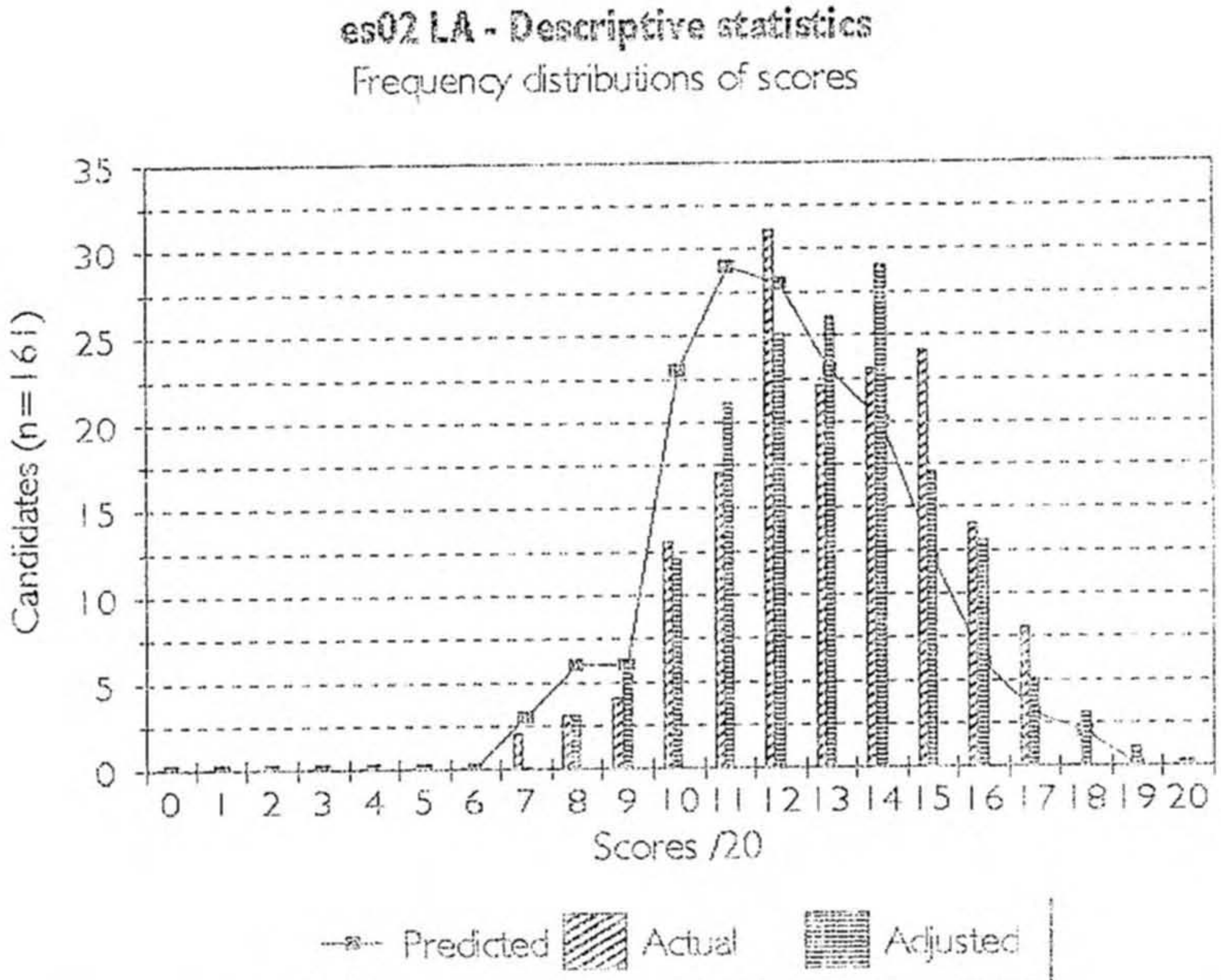
19. La adquisición de conocimientos en materia de derechos humanos contribuye a ...

A la eliminación de estereotipos.
B su aplicación a la vida diaria.
C el avance de la justicia social.
D la valoración de las interconexiones.
E la exploración de puntos de vista.

20. Cada acontecimiento ...

A es consecuencia de algo que ya se produjo y que ha influido en lo que se va a producir.
B es resultado de algo que no se produjo y que influirá en lo que se ha de producir.
C que se va a producir influyó en lo que ya se produjo.

D hunde sus raíces en algo que no se produjo, lo que influirá en lo que se va a producir.
E es resultado de algo que se produjo e influirá en lo que se ha de producir.


es02 LA - Descriptive statistics
Frequency distributions of scores
Candidates (n=161)
35
30
25
20
15
10
5
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Scores /20
Predicted
Actual
Adjusted

KEY.
PREDICTED

En la prueba es02, las alternativas designadas por los autores como las correctas eran las siguientes:

															N:
	1. ✓	**A**	B	C	D	E	F	G	H	I	J	K	L		I
	2. ✓	A	B	C	D	**E**	F	G	H	I	J	K	L		E
	3. ✓	A	B	C	D	E	F	G	**H**	I	J	K	L		AP
	4. ✓	A	B	C	D	E	F	G	H	I	**J**	K	L		BE
	5. ✓	A	B	C	D	E	F	G	H	I	J	**K**	L		IS
	6. ✓	V	**F**	?											BN
	7. ✓	V	**F**	?											CA
	8.	V	[✓ F] 0.45	**?** 0.43										Stet.	CF
	9. ✓	**V**	F	?											CH
	10. ✓	V	F	**?**											CN
0.[illegible]	11. ✓	A	B	C	D	**E**									CT
0.36	12.	[✓ A] 0.81	B	C	**D** 0.86	E								Exit	CY
0.75	13. ✓	A	**B**	C	D	E									DC
[illegible]	14.	A	[✓ B] 0.57	**C** 0.81	D	E								accept change	DE
0.75	15.	**A** 0.14	[✓ B] 0.78	C	D	E								Stet	DN
0.[illegible]	16.	[✓ A] 0.59	B	C	**D** 0.26	E								Stet.	DN
0.73	17. ✓	A	**B**	C	D	E									EA
0.40	18. ✓	A	B	C	D	**E**									EJ
0.75	19. ✓	A	B	**C**	D	E									E~~[illegible]~~N
0.77	20. ✓	A	B	C	D	**E**									EV.

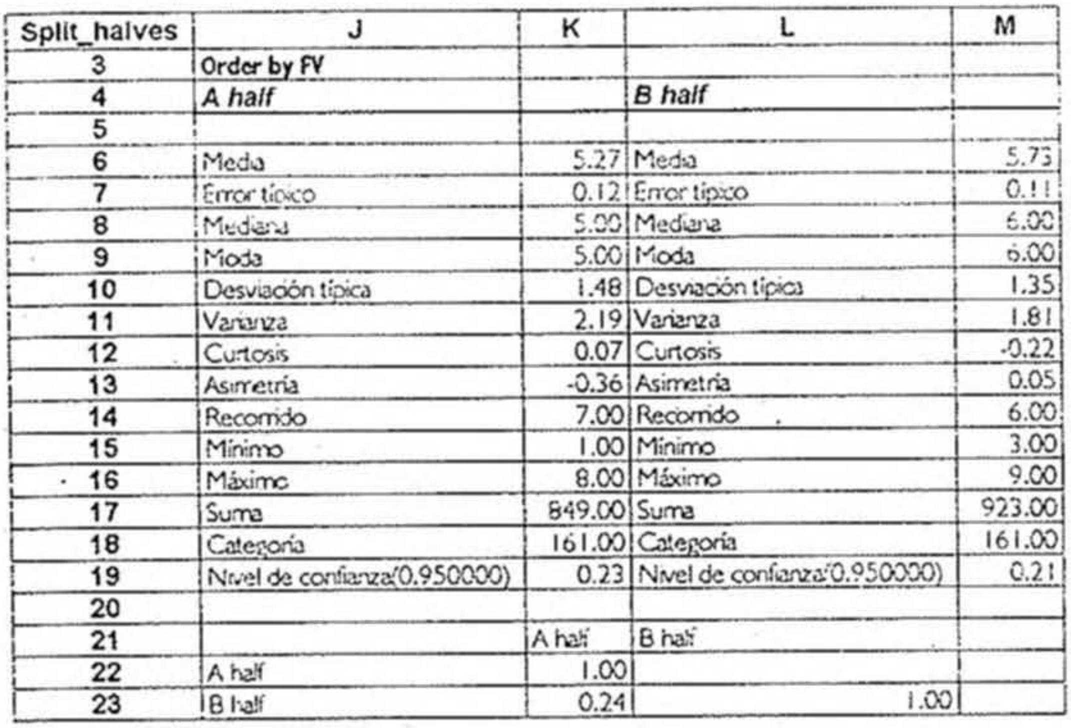
En la prueba es02, las alternativas que se adjudicaron ser las 'correctas' tras el análisis de los valores FV eran:

A:

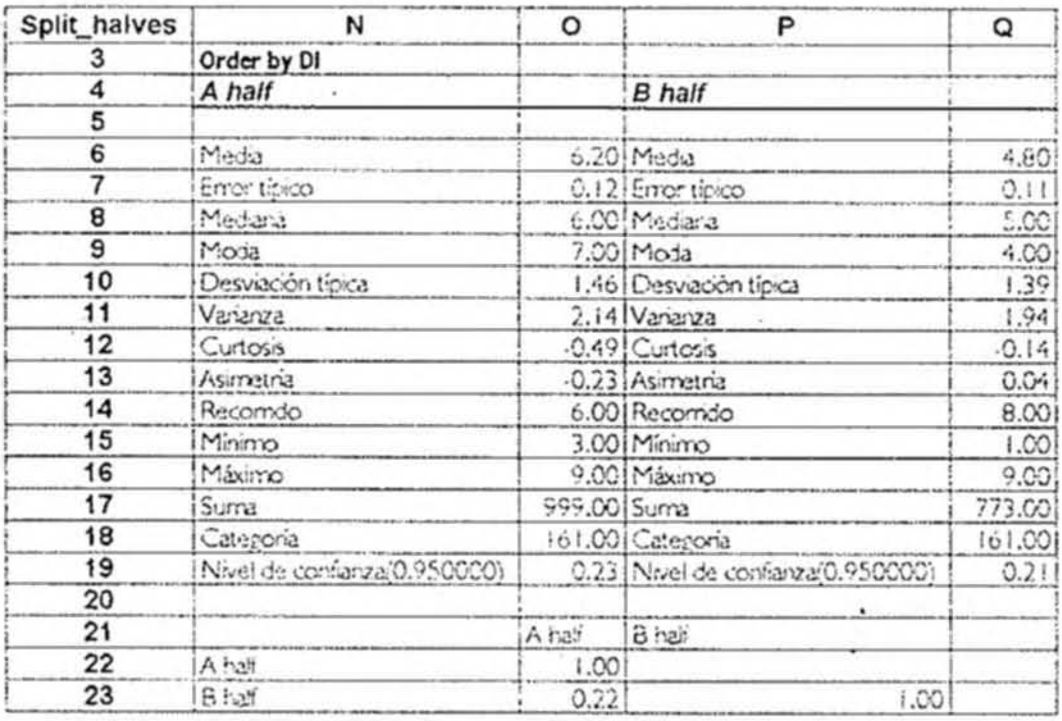
	1.	**A**	B	C	D	E	F	G	H	I	J	K	L	I
	2.	A	B	C	D	**E**	F	G	H	I	J	K	L	E
	3.	A	B	C	D	E	F	G	**H**	I	J	K	L	AP
	4.	A	B	C	D	E	F	G	H	I	**J**	K	L	BE
	5.	A	B	C	D	E	F	G	H	I	J	**K**	L	BS
	6.	V	**F**	?										BW
	7.	V	**F**	?										CA CF
	8.	V	F	**?**										CH
	9.	**V**	F	?										
	10.	V	F	**?**										CN
	11.	A	B	C	D	**E**								CT
X	12.	A	B	C	**D**	Ese elimina ya que hay dos alternativas correctas								~~G~~
	13.	A	**B**	C	D	E								DC
X	14.	A	**B**	C	D	E	CHANGE!							DI
	15.	**A**	B	C	D	E								DN
	16.	A	B	C	**D**	E								DW
	17.	A	**B**	C	D	E								EA
	18.	A	B	C	D	**E**								EJ
	19.	A	B	**C**	D	E								EN
	20.	A	B	C	D	**E**								EV

27/12/95

J	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	es02 Índice de discriminación																				
2																					
3																					
4		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
5	Correcto 'U'	44	44	32	44	38	36	35	29	42	28	41	0	38	32	11	15	25	27	32	44
6	Correcto 'L'	41	44	23	35	26	12	16	7	40	10	39	0	26	18	3	10	11	10	1	37
7	Nº 0 27,5%	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44	44
8	Índice D	0,07	0,00	0,20	0,20	0,27	0,54	0,43	0,50	0,05	0,41	0,05	0,00	0,27	0,32	0,18	0,11	0,32	0,38	0,70	0,16
9		✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓	✓				✓	✓
10																					
11																					
12																					
13									Q8						Q14	Q15	Q16				
14	Correcto 'U'								12						11	31	24				
15	Correcto 'L'								26						19	37	24				
16	Nº 0 27,5%								44						44	44	44				
17	Índice D								-0,32						-0,18	-0,14	0,00				



Split_halves	J	K	L	M
3	**Order by FV**			
4	***A half***		***B half***	
5				
6	Media	5.27	Media	5.73
7	Error típico	0.12	Error típico	0.11
8	Mediana	5.00	Mediana	6.00
9	Moda	5.00	Moda	6.00
10	Desviación típica	1.48	Desviación típica	1.35
11	Varianza	2.19	Varianza	1.81
12	Curtosis	0.07	Curtosis	-0.22
13	Asimetría	-0.36	Asimetría	0.05
14	Recorrido	7.00	Recorrido	6.00
15	Mínimo	1.00	Mínimo	3.00
16	Máximo	8.00	Máximo	9.00
17	Suma	849.00	Suma	923.00
18	Categoría	161.00	Categoría	161.00
19	Nivel de confianza(0.950000)	0.23	Nivel de confianza(0.950000)	0.21
20				
21		A half	B half	
22	A half	1.00		
23	B half	0.24	1.00	



Split_halves	N	O	P	Q
3	**Order by DI**			
4	*A half*		*B half*	
5				
6	Media	6.20	Media	4.80
7	Error típico	0.12	Error típico	0.11
8	Mediana	6.00	Mediana	5.00
9	Moda	7.00	Moda	4.00
10	Desviación típica	1.46	Desviación típica	1.39
11	Varianza	2.14	Varianza	1.94
12	Curtosis	-0.49	Curtosis	-0.14
13	Asimetría	-0.23	Asimetría	0.04
14	Recorrido	6.00	Recorrido	8.00
15	Mínimo	3.00	Mínimo	1.00
16	Máximo	9.00	Máximo	9.00
17	Suma	999.00	Suma	773.00
18	Categoría	161.00	Categoría	161.00
19	Nivel de confianza(0.950000)	0.23	Nivel de confianza(0.950000)	0.21
20				
21		A half	B half	
22	A half	1.00		
23	B half	0.22	1.00	

K	A	B	C	
1	es02 Summary table: FV and D			
2				
3				
4	Item	FV	D	
5	Q1	0.96	0.07	✗
6	Q2	1.00	0.00	✓
7	Q3	0.73	0.20	✗
8	Q4	0.90	0.20	✗
9	Q5	0.76	0.27	✗
10	Q6	0.57	0.54	✗
11	Q7	0.56	0.43	✗
12	Q8	0.43	0.50	✗
13	Q9	0.93	0.05	✗
14	Q10	0.47	0.41	✗
15	Q11	0.89	0.05	✗
16	Q12	0.00	0.00	
17	Q13	0.75	0.27	✗
18	Q14	0.57	0.32	✗
19	Q15	0.14	0.18	✗
20	Q16	0.26	0.11	✗
21	Q17	0.43	0.32	✗
22	Q18	0.38	0.38	✗
23	Q19	0.35	0.70	✗
24	Q20	0.94	0.16	✗

30/10/95. oh.

Stat.wb1 B:A1..M143

Valores FV - Prueba Julio 93 - Lengua A (ES02) - Archivo eso2~~u.wb1~~| Hoja B POBLACION: 161 CANDIDATOS.

B	A	B	C	D	E	F	G	H	I	J	K	L	M
1	pregunta nº	Q1A	Q1B	Q1C	Q1D	Q1E	Q1F	Q1G	Q1H	Q1I	Q1J	Q1K	Q1L
2	nº aciertos	154,0	1,0	2,0	4,0	0,0	0,0	0,0	0,0	1,0	0,0	0,0	0,
3	FV	✓0,96	0,01	0,01	0,02	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,0
4	nº respuestas	161,0											
5	respuestas fallidas	-0,0											
6	%	-0,0	A: I										
7													
8	pregunta nº	Q2A	Q2B	Q2C	Q2D	Q2E	Q2F	Q2G	Q2H	Q2I	Q2J	Q2K	Q2L
9	nº aciertos	0,0	1,0	0,0	1,0	161,0	1,0	0,0	0,0	0,0	0,0	0,0	0,
10	FV	0,00	0,01	0,00	0,01	1,00 ✓	0,01	0,00	0,00	0,00	0,00	0,00	0,0
11	nº respuestas	164,0											
12	respuestas fallidas	-3,0											
13	%	-1,9	A: 2										
14													
15	pregunta nº	Q3A	Q3B	Q3C	Q3D	Q3E	Q3F	Q3G	Q3H	Q3I	Q3J	Q3K	Q3L
16	nº aciertos	1,0	6,0	2,0	7,0	0,0	2,0	3,0	117,0	18,0	3,0	0,0	4,
17	FV	0,01	0,04	0,01	0,04	0,00	0,01	0,02	0,73 ✓	0,11	0,02	0,00	0,0
18	nº respuestas	163											
19	respuestas fallidas	-2,0											
20	%	-1,2	A: 1P										
21													
22	pregunta nº	Q4A	Q4B	Q4C	Q4D	Q4E	Q4F	Q4G	Q4H	Q4I	Q4J	Q4K	Q4L
23	nº aciertos	0,0	1,0	1,0	0,0	3,0	2,0	0,0	1,0	4,0	145,0	2,0	1.
24	FV	0,00	0,01	0,01	0,00	0,02	0,01	0,00	0,01	0,02	0,90 ✓	0,01	0,0
25	nº respuestas	160											
26	respuestas fallidas	1,0											
27	%	0,6	A: JE										
28													
29	pregunta nº	Q5A	Q5B	Q5C	Q5D	Q5E	Q5F	Q5G	Q5H	Q5I	Q5J	Q5K	Q5L
30	nº aciertos	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	4,0	0,0	122,0	36,
31	FV	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,00	0,78 ✓	0,2
32	nº respuestas	162,0											
33	respuestas fallidas	-1,0											
34	%	-0,6	A: 3S										

A15..M38

161 — 36%
288 — 64%
449

Valores FV - Prueba Julio 93 - Lengua A (ES02) - Archivo eso2fv.wb1 Hoja B

B	A	B	C	D	E	F		
35								A4C .. D,73
36	pregunta nº	Q6V	Q6F	Q6?			A: BW	
37	nº aciertos	66,0	92,0	2,0				
38	FV	0,41	0,57 ✓	0,01				
39	nº respuestas	160,0						
40	respuestas fallidas	1,0						
41	%	0,6						
42								
43	pregunta nº	Q7V	Q7F	Q7?			A: CA	
44	nº aciertos	57,0	90,0	14,0				
45	FV	0,35	0,56 ✓	0,09				
46	nº respuestas	161,0						
47	respuestas fallidas	0,0						
48	%	0,0						
49								
50	pregunta nº	Q8V	Q8F	Q8?			A: CE	
51	nº aciertos	19,0	72,0	70,0				
52	FV	0,12	0,45	0,43 ✓				
53	nº respuestas	161,0						
54	respuestas fallidas	0,0						
55	%	0,0						
56								
57	pregunta nº	Q9V	Q9F	Q9?			A: CH	
58	nº aciertos	150,0	9,0	2,0				
59	FV	✓ 0,93	0,06	0,01				
60	nº respuestas	161,0						
61	respuestas fallidas	0,0						
62	%	0,0						
63								
64	pregunta nº	Q10V	Q10F	Q10?			A: CN	
65	nº aciertos	40,0	46,0	75,0				
66	FV	0,25	0,29	0,47 ✓				
67	nº respuestas	161,0						
68	respuestas fallidas	0,0						
69	%	0,0						
70								
71	pregunta nº	Q11A	Q11B	Q11C	Q11D	Q11E	A: CT.	A75 .. F 143
72	nº aciertos	0,0	0,0	0,0	18,0	143,0		
73	FV	0,00	0,00	0,00	0,11	0,89 ✓		
74	nº respuestas	161,0						
75	respuestas fallidas	0,0						
76	%	0,0						
77								
78	pregunta nº	Q12A ✓	Q12B	Q12C	Q12D ✓	Q12E	∅ out	
79	nº aciertos	58,0	5,0	27,0	58,0	12,0		
80	FV	0,36	0,03	0,17	0,36	0,07		
81	nº respuestas	160,0						
82	respuestas fallidas	1,0						
83	%	0,6						
84								
85	pregunta nº	Q13A	Q13B	Q13C	Q13D	Q13E	A: DC	
86	nº aciertos	30,0	121,0	2,0	3,0	4,0		
87	FV	0,19	0,75 ✓	0,01	0,02	0,02		
88	nº respuestas	160,0						
89	respuestas fallidas	1,0						
90	%	0,6						

Valores FV - Prueba Julio 93 - Lengua A (ES02) - Archivo eso2fv.wb1 Hoja B

B	A	B	C	D	E	F	
91							
92	pregunta n°	Q14A	Q14B	Q14C	Q14D	Q14E	A: DI
93	n° aciertos	14,0	91,0	50,0	6,0	2,0	
94	FV	0,09	0,57 ✓	0,31	0,04	0,01	
95	n° respuestas	163,0					
96	respuestas fallidas	-2,0					
97	%	-1,2					
98							
99	pregunta n°	Q15A	Q15B	Q15C	Q15D	Q15E	A: DO
100	n° aciertos	22,0	125,0	7,0	1,0	4,0	
101	FV	0,14 ✓	0,78	0,04	0,01	0,02	
102	n° respuestas	159,0					
103	respuestas fallidas	2,0					
104	%	1,2					
105							
106	pregunta n°	Q16A	Q16B	Q16C	Q16D	Q16E	A: DT
107	n° aciertos	95,0	15,0	6,0	42,0	1,0	
108	FV	0,59	0,09	0,04	0,26 ✓	0,01	
109	n° respuestas	159,0					
110	respuestas fallidas	2,0					
111	%	1,2					
112							
113	pregunta n°	Q17A	Q17B	Q17C	Q17D	Q17E	A: EA
114	n° aciertos	6,0	70,0	32,0	35,0	18,0	
115	FV	0,04	0,43 ✓	0,20	0,22	0,11	
116	n° respuestas	161,0					
117	respuestas fallidas	0,0					
118	%	0,0					
119							
120	pregunta n°	Q18A	Q18B	Q18C	Q18D	Q18E	A: EJ
121	n° aciertos	50,0	1,0	27,0	18,0	66,0	
122	FV	0,31	0,01	0,17	0,11	0,38	
123	n° respuestas	157,0					
124	respuestas fallidas	4,0					
125	%	0,0					
126							
127	pregunta n°	Q19A	Q19B	Q19C	Q19D	Q19E	A: EN
128	n° aciertos	52,0	8,0	56,0	23,0	21,0	
129	FV	0,32	0,05	0,35 ✓	0,14	0,13	
130	n° respuestas	160,0					
131	respuestas fallidas	1,0					
132	%	0,6					
133							
134	pregunta n°	Q20A	Q20B	Q20C	Q20D	Q20E	A. EV
135	n° aciertos	3,0	3,0	2,0	2,0	151,0	
136	FV	0,02	0,02	0,01	0,01	0,94 ✓	
137	n° respuestas	161,0					
138	respuestas fallidas	0,0					
139	%	0,0					

Sub-test in02

Today's royal bride: subversive modern or anti-feminist reactionary?

Japan marries self-sacrifice with nostalgia

BY IAN BURUMA

A MODERN Japanese monarchs have liked to copy the style of the British royal family. The late Emperor Hirohito, in particular, admired the casual, golf-playing dash of Edward VIII. He was perhaps a rather odd role model but, his abdication and other problems notwithstanding, British royalty remained until recently the *ne plus ultra* of monarchical grandeur.

B But Crown Prince Naruhito and Masako Owada marry today against a background of changed attitudes to the British monarchy. Now, it is viewed by Japanese royalists as a parable of failure, an example of how things can go terribly wrong. Look, for example, at the difference between the Princess of Wales and Miss Owada, a commoner, whose wedding to Prince Naruhito today will make her crown princess.

Light-hearted: Masako Owada, left, and her sister Meiko, with a paper lantern given by well-wishers

C The Princess of Wales, one is told (and not only by Japanese conservatives), is the perfect example of modern selfishness. By refusing to play the game, by insisting on her freedom, she has jeopardised an ancient institution. How noble, in contrast, seems Miss Owada: after refusing for seven years to give up a promising diplomatic career for the gilded prison of the Japanese imperial court, she was prevailed on to put aside "selfish" concerns and marry a man for whom she has not expressed any affection. That it was an act of self-sacrifice is clear from her language. While once she said that she would never give up her freedom, she now promises to make herself "useful to the imperial household". Self-sacrifice, for the country, for a feudal lord, for a business corporation, for a husband, is the greatest traditional virtue in a Confucian society, or indeed in any society where the strong can impose virtues on the weak. It is a virtue which fills some Western reactionaries with nostalgia.

D Such nostalgics like to project the shortcomings of their own societies (the selfish, materialistic West) on to other, faraway places (the disciplined, spiritual East). But Miss Owada's virtue is oddly out of step with trends in Japan too.

E Her arranged marriage comes when, for the first time in Japanese history, women are doing what the Princess of Wales has done: claiming the right to disentangle themselves from intolerable marriages. More Japanese women than men now ask for divorces. An increasing number of women are more interested in careers and economic independence than in traditional family life. Some (known derisively as "yellow cabs" — you can always get a ride) would rather travel abroad for casual sex than enter into subservient relationships with Japanese men. One could call this selfish and sad. One could also say that society has not kept pace with its most intelligent and independent women.

F Although there was never any hint of the yellow cab about her, Miss Owada was such a woman. This may be partly why Prince Naruhito was attracted to her, rather than to some simpering blueblood who played by all the old rules. This is what makes the marriage so interesting: either Miss Owada's sacrifice is a reactionary blow against female independence, or she is a subversive modern in the heart of the nation's still-sacred institution.

Ian Buruma is the author of A Japanese Mirror.

English - Test of Reading Skills - Text 2

1st SURNAME:

2nd SURNAME:

NAME:

DNI:

TOTAL:

Underline the appropriate letter or symbol.

1. A B C D E F

2. A B C D E F

3. A B C D E F

4. A B C D E F

5. A B C D E F

6. T F ?

7. T F ?

8. T F ?

9. T F ?

10. T F ?

11. A B C D E

12. A B C D E

13. A B C D E

14. A B C D E

15. A B C D E

16. A B C D E

17. A B C D E

18. A B C D E

19. A B C D E

20. A B C D E

English - Test of Reading Skills - Text 2

"Japan marries self-sacrifice with nostalgia"

Scan the text and identify for which paragraph each of these sentences could be the title.

1. A distant admirer.

2. Contrasting attitudes.

3. Out of step with society?

4. The enigma of Miss Owada's decision.

5. A view from the present.

Choose T (True), F (False), or ? ("there is no evidence in the passage to support this").

6. The Japanese see the Princess of Wales as an example of modern social change.

7. Miss Owada decided to accept her arranged marriage out of a sense of tradition.

8. Some Western observers would wish to impose Confucian traditions on their own societies.

9. Prince Naruhito has spoken out against the attitudes displayed by the British Royal Family.

10. The stability of Japanese family life has been shaken by the so-called "yellow cabs".

Choose the most appropriate alternative and underline the letter on the answer sheet.

11. Miss Owada's decision represents ...

    A a clear change of mind.
    B the fulfilment of her dreams.
    C an acquiescence to family pressure.
    D her acceptance of destiny.
    E her desire to be subversive.

12. Edward VIII was admired by the Japanese because he ...

    A abdicated the throne.
    B maintained traditions.
    C epitomised royalty.
    D exhibited eccentricities.
    E showed a certain style.

13. Once she becomes a member of the Japanese imperial court Miss Owada will probably ...

    A be in a position to bring about change.
    B represent female independence in modern Japan.
    C be referred to as being a "yellow cab".
    D find her life comfortable, but limited.
    E follow the example set by the Princess of Wales.

14. Before deciding to marry Crown Prince Naruhito, Miss Owada was ...

    A an anti-feminist.
    B a traditional Confucian.
    C a model of selfishness.
    D a "yellow cab".
    E a career woman.

15. The position held by the monarchy in Japanese society appears to be ...

    A in jeopardy.
    B under review.
    C unquestioned.
    D intolerable.
    E justified.

16. The changes in modern-day Japanese society are ...

    A creating concern amongst traditionalists.
    B a cause for nostalgia in both West and East.
    C slowly being undermined by subversives.
    D no longer as clear as they used to be.
    E clearly seen in the life of the royal family.

17. In paragraph E, from the sentences "One could call this ... independent women." we can assume that the author of the article does himself believe ...

    A that this is "selfish and sad".
    B that "society has not kept pace".
    C neither 'A' nor 'B' is true.
    D both 'A' and 'B' are true.
    E something completely different.

18. In paragraph B, the expression "a parable of failure" refers to ...

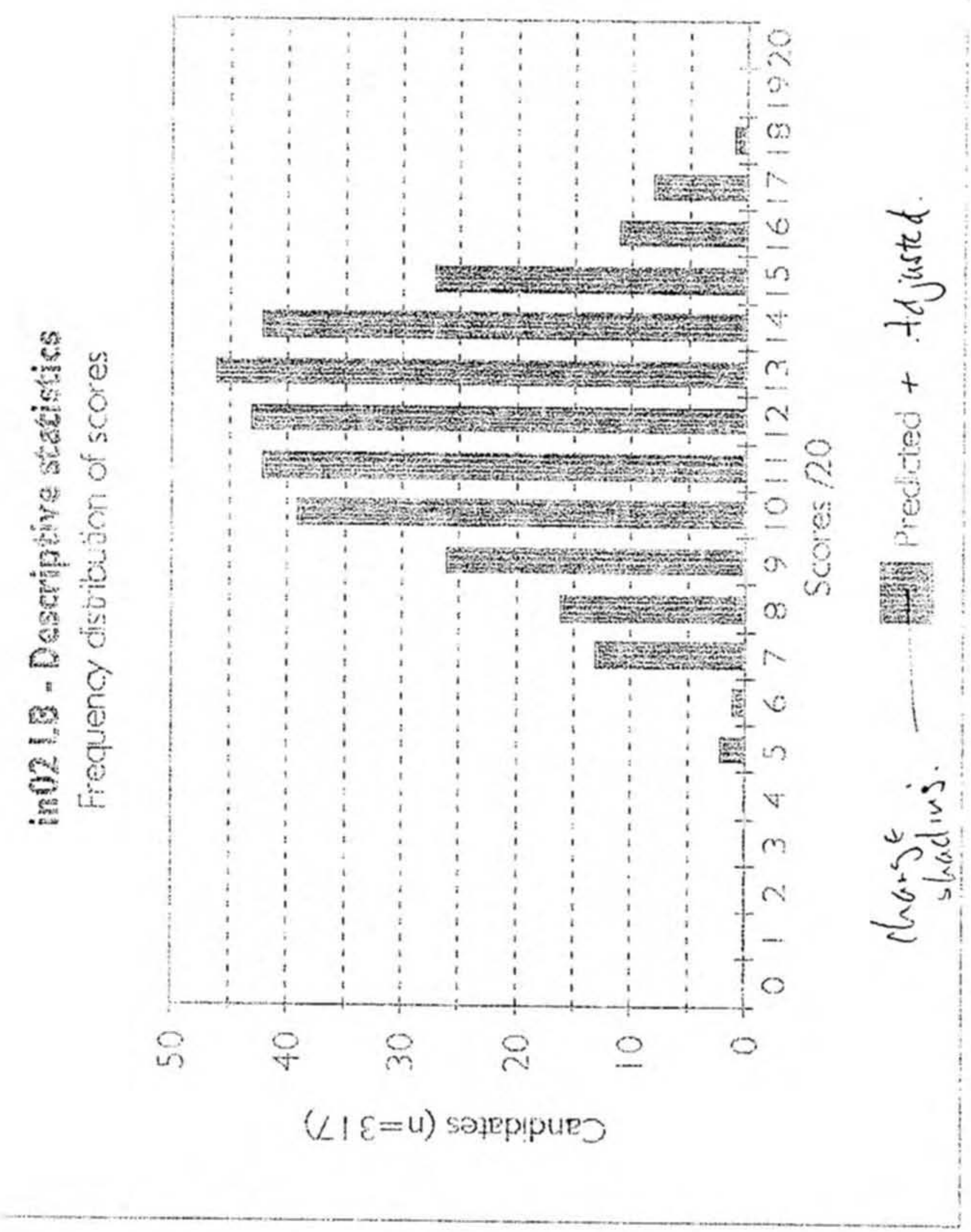
    A Edward VIII's decision to abdicate the throne.
    B Japanese royalists' views on the monarchy.
    C the marriage of the Prince and Princess of Wales.
    D the marriage of Crown Prince Naruhito to Miss Owada.
    E the current situation of the British Royal Family.

19. In paragraph C, the phrase "to play the game" is synonymous with ...

    A victory.
    B conformity.
    C defeat.
    D stalemate.
    E rebellion.

20. In the phrase "That it was an act of self-sacrifice ..." in paragraph C, the word "it" refers to ...

    A being selfish.
    B playing the game.
    C getting married.
    D being useful.
    E pursuing a career.


in02 I.B - Descriptive statistics
Frequency distribution of scores
Candidates (n=317)
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Scores /20
Predicted + Adjusted.
change shading.

PREDICTED

+

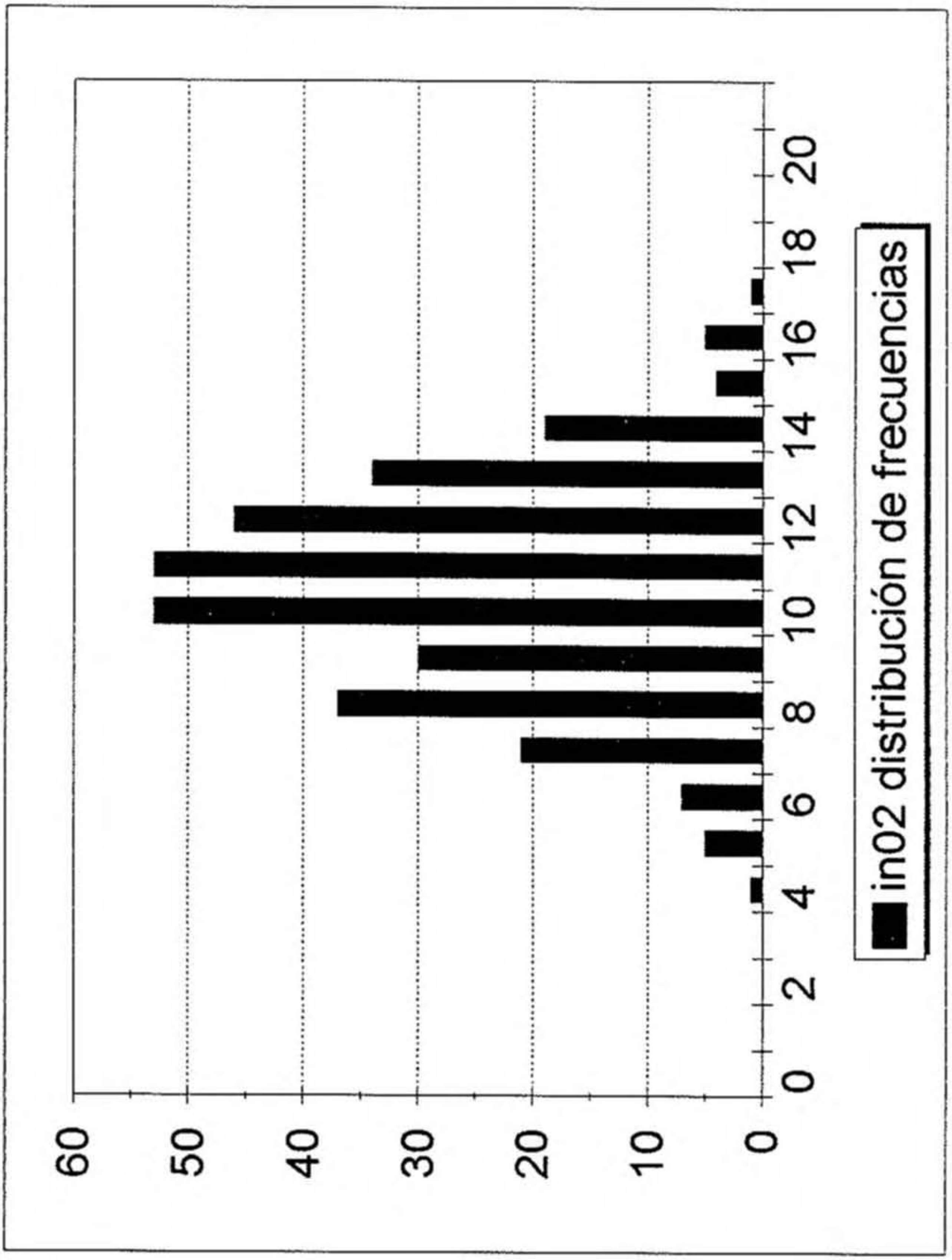
ADJUSTED

En la prueba in02, las alternativas designadas por los autores como las correctas eran las que figuran aquí abajo. Después del análisis de los resultados actuales de la prueba se ha decidido mantener como correctas estas mismas alternativas:

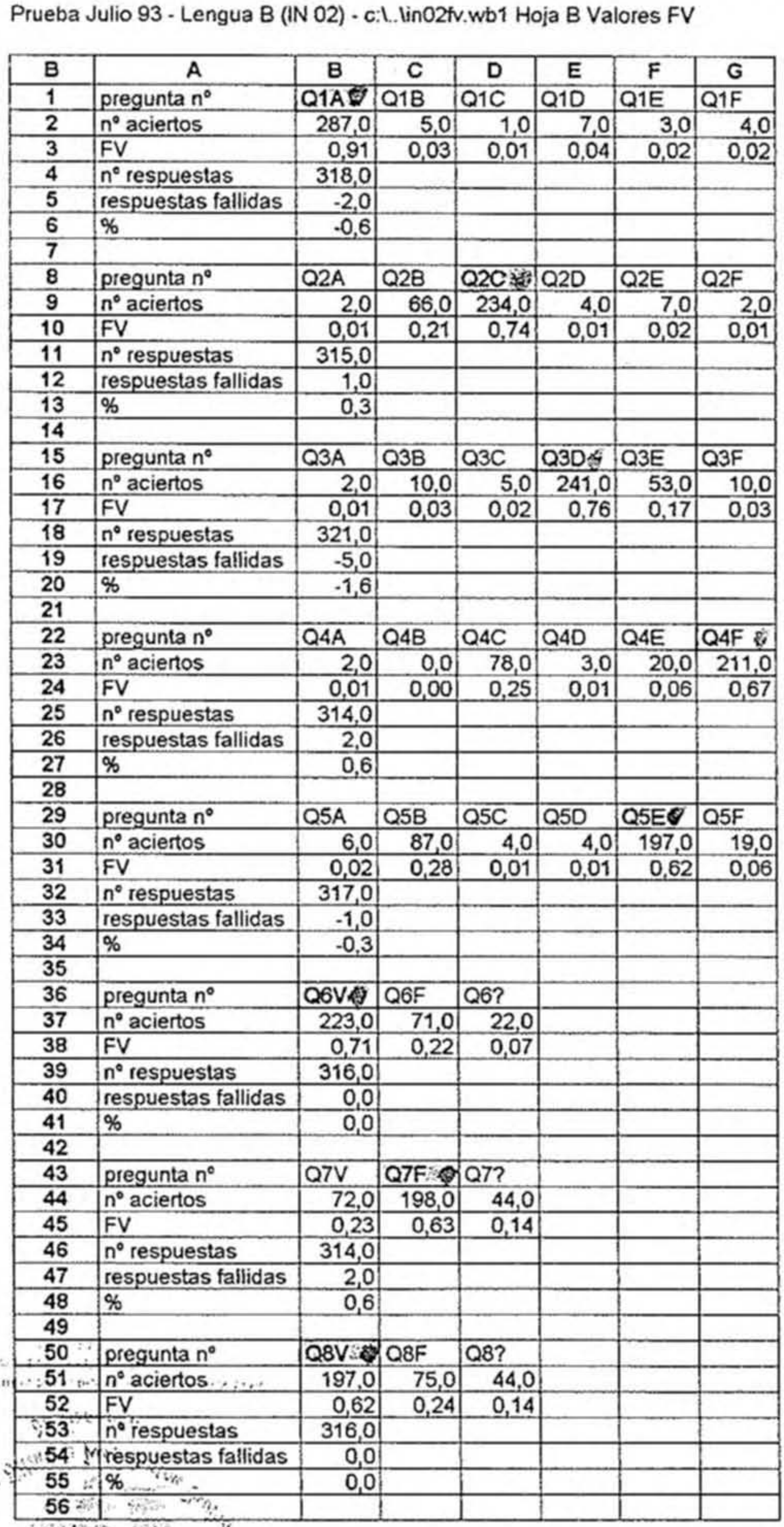
1.	***A***	B	C	D	E	F
2.	A	B	***C***	D	E	F
3.	A	B	C	***D***	E	F
4.	A	B	C	D	E	***F***
5.	A	***B***	C	D	E	F
6.	***V***	F	?			
7.	V	F	***?***			
8.	***V***	F	?			
9.	V	F	***?***			
10.	V	F	***?***			
11.	***A***	B	C	D	E	
12.	A	B	C	D	***E***	
13.	A	B	C	***D***	E	
14.	A	B	C	D	***E***	
15.	A	B	***C***	D	E	
16.	***A***	B	C	D	E	
17.	A	***B***	C	D	E	
18.	A	B	C	D	***E***	
19.	A	***B***	C	D	E	
20.	A	B	***C***	D	E	

Shade squares?

Printout Actual.



c:\document\thesis\results\stats\in02\in02stat.wb1



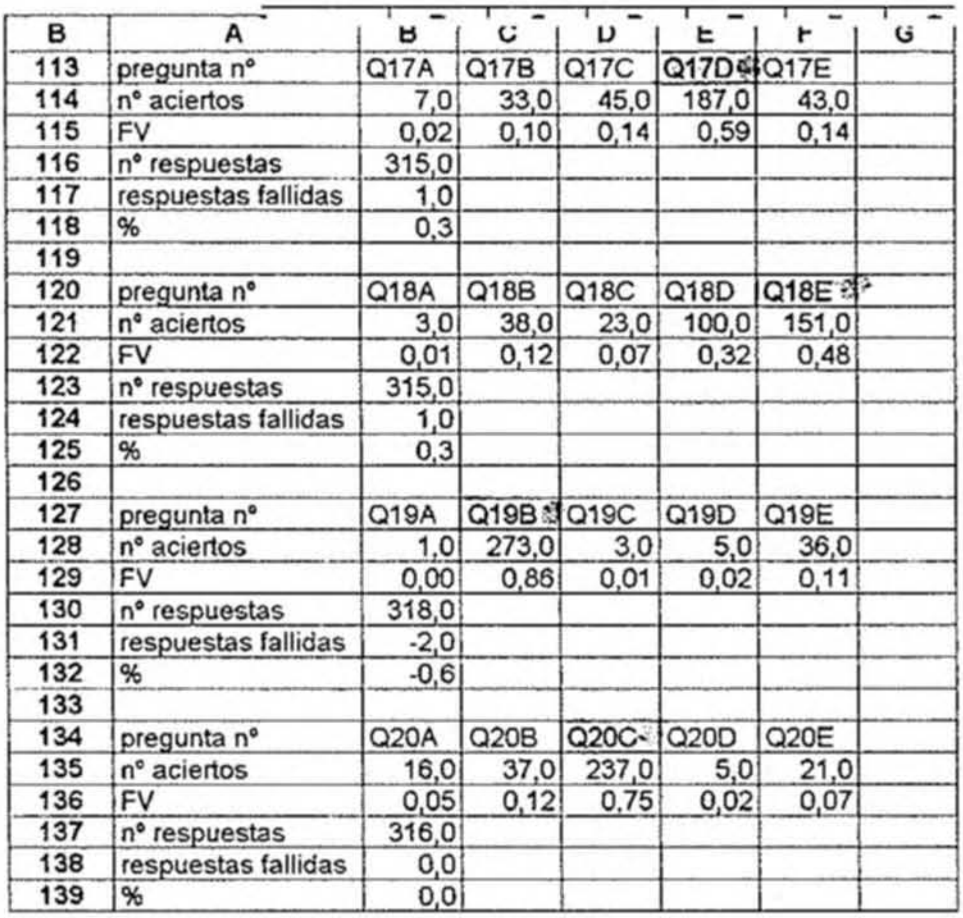
Prueba Julio 93 - Lengua B (IN 02) - c:\..\in02fv.wb1 Hoja B Valores FV

B	A	B	C	D	E	F	G
1	pregunta nº	Q1A✓	Q1B	Q1C	Q1D	Q1E	Q1F
2	nº aciertos	287,0	5,0	1,0	7,0	3,0	4,0
3	FV	0,91	0,03	0,01	0,04	0,02	0,02
4	nº respuestas	318,0					
5	respuestas fallidas	-2,0					
6	%	-0,6					
7							
8	pregunta nº	Q2A	Q2B	Q2C✓	Q2D	Q2E	Q2F
9	nº aciertos	2,0	66,0	234,0	4,0	7,0	2,0
10	FV	0,01	0,21	0,74	0,01	0,02	0,01
11	nº respuestas	315,0					
12	respuestas fallidas	1,0					
13	%	0,3					
14							
15	pregunta nº	Q3A	Q3B	Q3C	Q3D✓	Q3E	Q3F
16	nº aciertos	2,0	10,0	5,0	241,0	53,0	10,0
17	FV	0,01	0,03	0,02	0,76	0,17	0,03
18	nº respuestas	321,0					
19	respuestas fallidas	-5,0					
20	%	-1,6					
21							
22	pregunta nº	Q4A	Q4B	Q4C	Q4D	Q4E	Q4F ✓
23	nº aciertos	2,0	0,0	78,0	3,0	20,0	211,0
24	FV	0,01	0,00	0,25	0,01	0,06	0,67
25	nº respuestas	314,0					
26	respuestas fallidas	2,0					
27	%	0,6					
28							
29	pregunta nº	Q5A	Q5B	Q5C	Q5D	Q5E✓	Q5F
30	nº aciertos	6,0	87,0	4,0	4,0	197,0	19,0
31	FV	0,02	0,28	0,01	0,01	0,62	0,06
32	nº respuestas	317,0					
33	respuestas fallidas	-1,0					
34	%	-0,3					
35							
36	pregunta nº	Q6V✓	Q6F	Q6?			
37	nº aciertos	223,0	71,0	22,0			
38	FV	0,71	0,22	0,07			
39	nº respuestas	316,0					
40	respuestas fallidas	0,0					
41	%	0,0					
42							
43	pregunta nº	Q7V	Q7F✓	Q7?			
44	nº aciertos	72,0	198,0	44,0			
45	FV	0,23	0,63	0,14			
46	nº respuestas	314,0					
47	respuestas fallidas	2,0					
48	%	0,6					
49							
50	pregunta nº	Q8V✓	Q8F	Q8?			
51	nº aciertos	197,0	75,0	44,0			
52	FV	0,62	0,24	0,14			
53	nº respuestas	316,0					
54	respuestas fallidas	0,0					
55	%	0,0					
56							

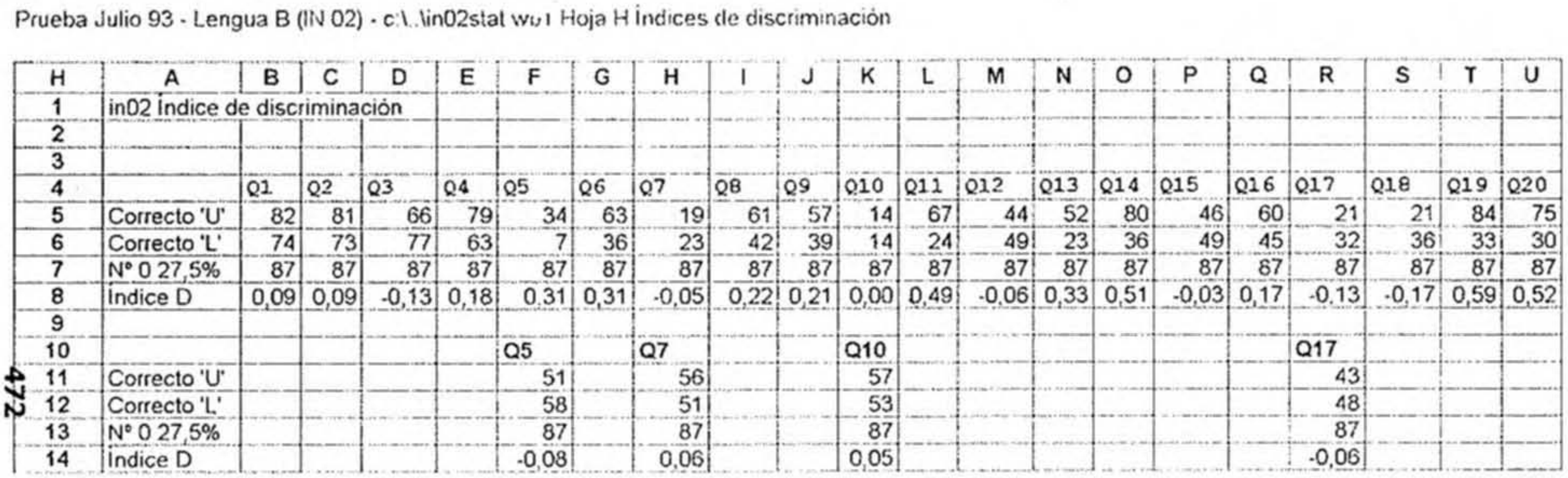
Prueba Julio 93 - Lengua B (IN 02) - c:\..\in02fv.wb1 Hoja B Valores FV

B	A	B	C	D	E	F	G
57	pregunta nº	Q9V	Q9F	Q9? ✓			
58	nº aciertos	44,0	95,0	176,0			
59	FV	0,14	0,30	0,56			
60	nº respuestas	315,0					
61	respuestas fallidas	1,0					
62	%	0,3					
63							
64	pregunta nº	Q10V ✓	Q10F	Q10?			
65	nº aciertos	202,0	76,0	38,0			
66	FV	0,64	0,24	0,12			
67	nº respuestas	316,0					
68	respuestas fallidas	0,0					
69	%	0,0					
70							
71	pregunta nº	Q11A ✓	Q11B	Q11C	Q11D	Q11E	
72	nº aciertos	163,0	7,0	28,0	61,0	62,0	
73	FV	0,52	0,02	0,09	0,19	0,20	
74	nº respuestas	321,0					
75	respuestas fallidas	-5,0					
76	%	-1,6					
77							
78	pregunta nº	Q12A	Q12B	Q12C	Q12D	Q12E ✓	
79	nº aciertos	63,0	71,0	41,0	28,0	113,0	
80	FV	0,20	0,22	0,13	0,09	0,36	
81	nº respuestas	316,0					
82	respuestas fallidas	0,0					
83	%	0,0					
84							
85	pregunta nº	Q13A	Q13B	Q13C	Q13D ✓	Q13E	
86	nº aciertos	109,0	40,0	9,0	129,0	26,0	
87	FV	0,34	0,13	0,03	0,41	0,08	
88	nº respuestas	313,0					
89	respuestas fallidas	3,0					
90	%	0,9					
91							
92	pregunta nº	Q14A	Q14B	Q14C	Q14D	Q14E ✓	
93	nº aciertos	1,0	12,0	42,0	28,0	232,0	
94	FV	0,00	0,04	0,13	0,09	0,73	
95	nº respuestas	315,0					
96	respuestas fallidas	1,0					
97	%	0,3					
98							
99	pregunta nº	Q15A	Q15B	Q15C ✓	Q15D	Q15E	
100	nº aciertos	52,0	74,0	99,0	44,0	43,0	
101	FV	0,16	0,23	0,31	0,14	0,14	
102	nº respuestas	312,0					
103	respuestas fallidas	4,0					
104	%	1,3					
105							
106	pregunta nº	Q16A ✓	Q16B	Q16C	Q16D	Q16E	
107	nº aciertos	146,0	63,0	32,0	27,0	43,0	
108	FV	0,46	0,20	0,10	0,09	0,14	
109	nº respuestas	311,0					
110	respuestas fallidas	5,0					
111	%	1,6					
112							

Prueba Julio 93 - Lengua B (IN 02) - c:\..\in02fv.wb1 Hoja B Valores FV



B	A	B	C	D	E	F	G
113	pregunta nº	Q17A	Q17B	Q17C	Q17D*	Q17E	
114	nº aciertos	7,0	33,0	45,0	187,0	43,0	
115	FV	0,02	0,10	0,14	0,59	0,14	
116	nº respuestas	315,0					
117	respuestas fallidas	1,0					
118	%	0,3					
119							
120	pregunta nº	Q18A	Q18B	Q18C	Q18D	Q18E*	
121	nº aciertos	3,0	38,0	23,0	100,0	151,0	
122	FV	0,01	0,12	0,07	0,32	0,48	
123	nº respuestas	315,0					
124	respuestas fallidas	1,0					
125	%	0,3					
126							
127	pregunta nº	Q19A	Q19B*	Q19C	Q19D	Q19E	
128	nº aciertos	1,0	273,0	3,0	5,0	36,0	
129	FV	0,00	0,86	0,01	0,02	0,11	
130	nº respuestas	318,0					
131	respuestas fallidas	-2,0					
132	%	-0,6					
133							
134	pregunta nº	Q20A	Q20B	Q20C*	Q20D	Q20E	
135	nº aciertos	16,0	37,0	237,0	5,0	21,0	
136	FV	0,05	0,12	0,75	0,02	0,07	
137	nº respuestas	316,0					
138	respuestas fallidas	0,0					
139	%	0,0					



Prueba Julio 93 - Lengua B (IN 02) - c:\..\in02stat wо1 Hoja H Indices de discriminación

H	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	in02 Indice de discriminación																				
2																					
3																					
4		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
5	Correcto 'U'	82	81	66	79	34	63	19	61	57	14	67	44	52	80	46	60	21	21	84	75
6	Correcto 'L'	74	73	77	63	7	36	23	42	39	14	24	49	23	36	49	45	32	36	33	30
7	Nº 0 27,5%	87	87	87	87	87	87	87	87	87	87	87	87	87	87	87	67	87	87	87	87
8	Indice D	0,09	0,09	-0,13	0,18	0,31	0,31	-0,05	0,22	0,21	0,00	0,49	-0,06	0,33	0,51	-0,03	0,17	-0,13	-0,17	0,59	0,52
9																					
10						Q5		Q7			Q10							Q17			
11	Correcto 'U'					51		56			57							43			
12	Correcto 'L'					58		51			53							48			
13	Nº 0 27,5%					87		87			87							87			
14	Indice D					-0,08		0,06			0,05							-0,06			

Prueba Julio 93 - Lengua B (IN 02) - c:\..\in02stat.wb1 Hoja D Estadística descriptiva

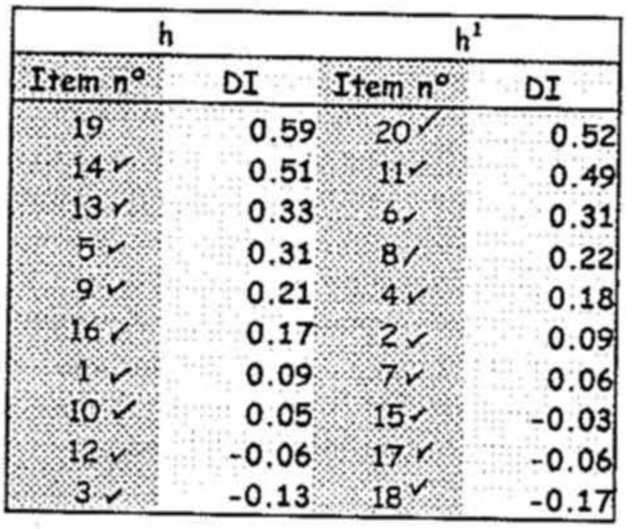
D	A	B	C	D
1	in02 estadística descriptiva			
2				
3				
4				
5	*Predicted*		*Actual*	
6				
7	Media	10,49	Media	12,33
8	Error típico	0,13	Error típico	0,14
9	Mediana	11,00	Mediana	12,00
10	Moda	10,00	Moda	12,00
11	Desviación típica	2,35	Desviación típica	2,50
12	Varianza	5,53	Varianza	6,27
13	Curtosis	-0,23	Curtosis	-0,45
14	Asimetría	-0,07	Asimetría	-0,22
15	Recorrido	13,00	Recorrido	13,00
16	Mínimo	4,00	Mínimo	5,00
17	Máximo	17,00	Máximo	18,00
18	Suma	3.314,00	Suma	3.896,00
19	Categoría	316,00	Categoría	316,00
20	Nivel de confianza(0,990000)	0,34	Nivel de confianza(0,990000)	0,36

Split halves

h		h^1	
Item n°	FV	Item n°	FV
17B ✓	0.12	10?	0.12
7? ✓	0.14	5B	0.28
15C ✓	0.31	12E	0.36
13D ✓	0.41	16A	0.46
18E ✓	0.48	11A	0.52
9? ✓	0.56	8V	0.63
4F ✓	0.67	6V	0.71
14E ✓	0.74	2C	0.74
20C ✓	0.75	3D	0.77
19B ✓	0.87	1A	0.91

Descriptive statistics

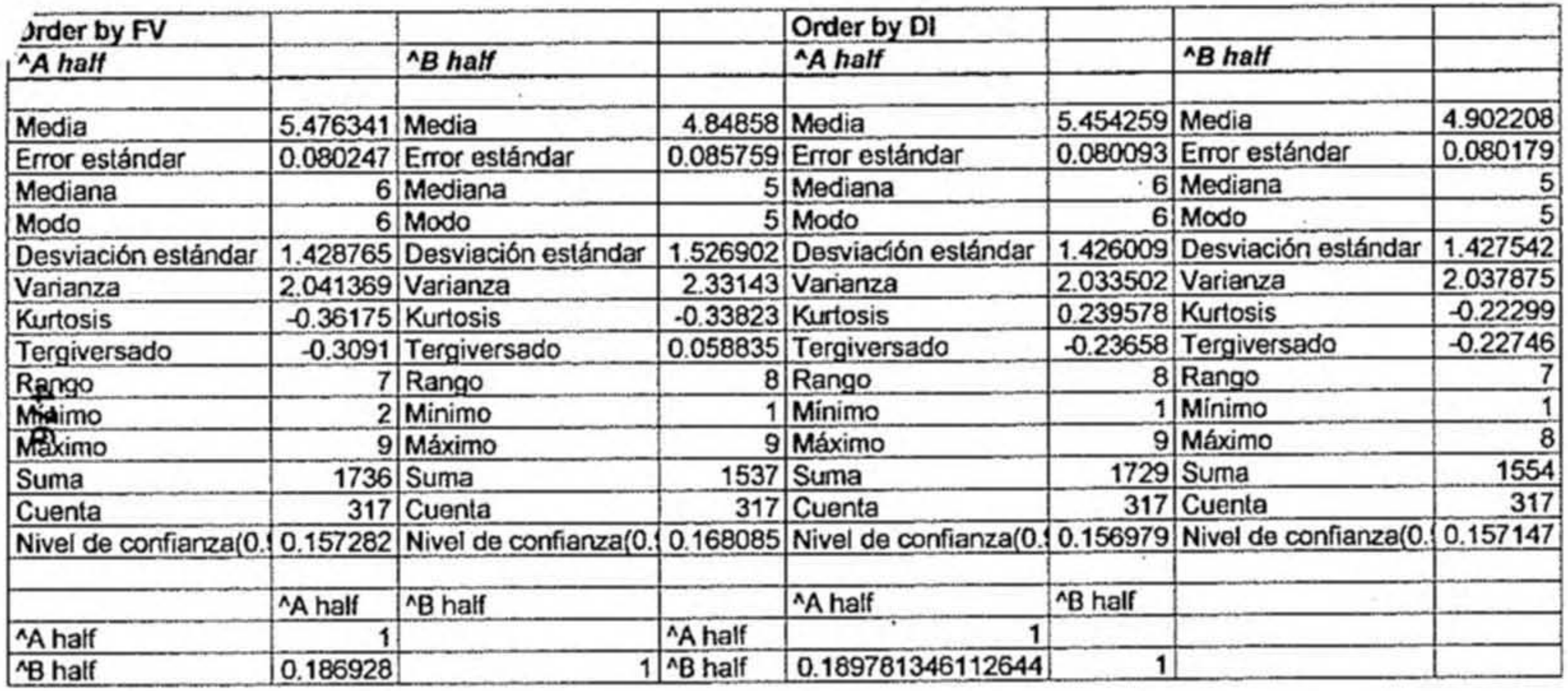
Split halves



h		h^1	
Item n°	DI	Item n°	DI
19	0.59	20 ✓	0.52
14 ✓	0.51	11 ✓	0.49
13 ✓	0.33	6 ✓	0.31
5 ✓	0.31	8 ✓	0.22
9 ✓	0.21	4 ✓	0.18
16 ✓	0.17	2 ✓	0.09
1 ✓	0.09	7 ✓	0.06
10 ✓	0.05	15 ✓	-0.03
12 ✓	-0.06	17 ✓	-0.06
3 ✓	-0.13	18 ✓	-0.17

Descriptive statistics

In∅2



Order by FV				Order by DI			
^A half		*^B half*		*^A half*		*^B half*	
Media	5.476341	Media	4.84858	Media	5.454259	Media	4.902208
Error estándar	0.080247	Error estándar	0.085759	Error estándar	0.080093	Error estándar	0.080179
Mediana	6	Mediana	5	Mediana	6	Mediana	5
Modo	6	Modo	5	Modo	6	Modo	5
Desviación estándar	1.428765	Desviación estándar	1.526902	Desviación estándar	1.426009	Desviación estándar	1.427542
Varianza	2.041369	Varianza	2.33143	Varianza	2.033502	Varianza	2.037875
Kurtosis	-0.36175	Kurtosis	-0.33823	Kurtosis	0.239578	Kurtosis	-0.22299
Tergiversado	-0.3091	Tergiversado	0.058835	Tergiversado	-0.23658	Tergiversado	-0.22746
Rango	7	Rango	8	Rango	8	Rango	7
Mínimo	2	Minimo	1	Mínimo	1	Mínimo	1
Máximo	9	Máximo	9	Máximo	9	Máximo	8
Suma	1736	Suma	1537	Suma	1729	Suma	1554
Cuenta	317	Cuenta	317	Cuenta	317	Cuenta	317
Nivel de confianza(0.!	0.157282	Nivel de confianza(0.!	0.168085	Nivel de confianza(0.!	0.156979	Nivel de confianza(0.!	0.157147
	^A half	^B half		^A half	^B half		
^A half	1		^A half	1			
^B half	0.186928	1	^B half	0.189781346112644	1		

~~10/02/07~~

Rank-order by Discrimination index

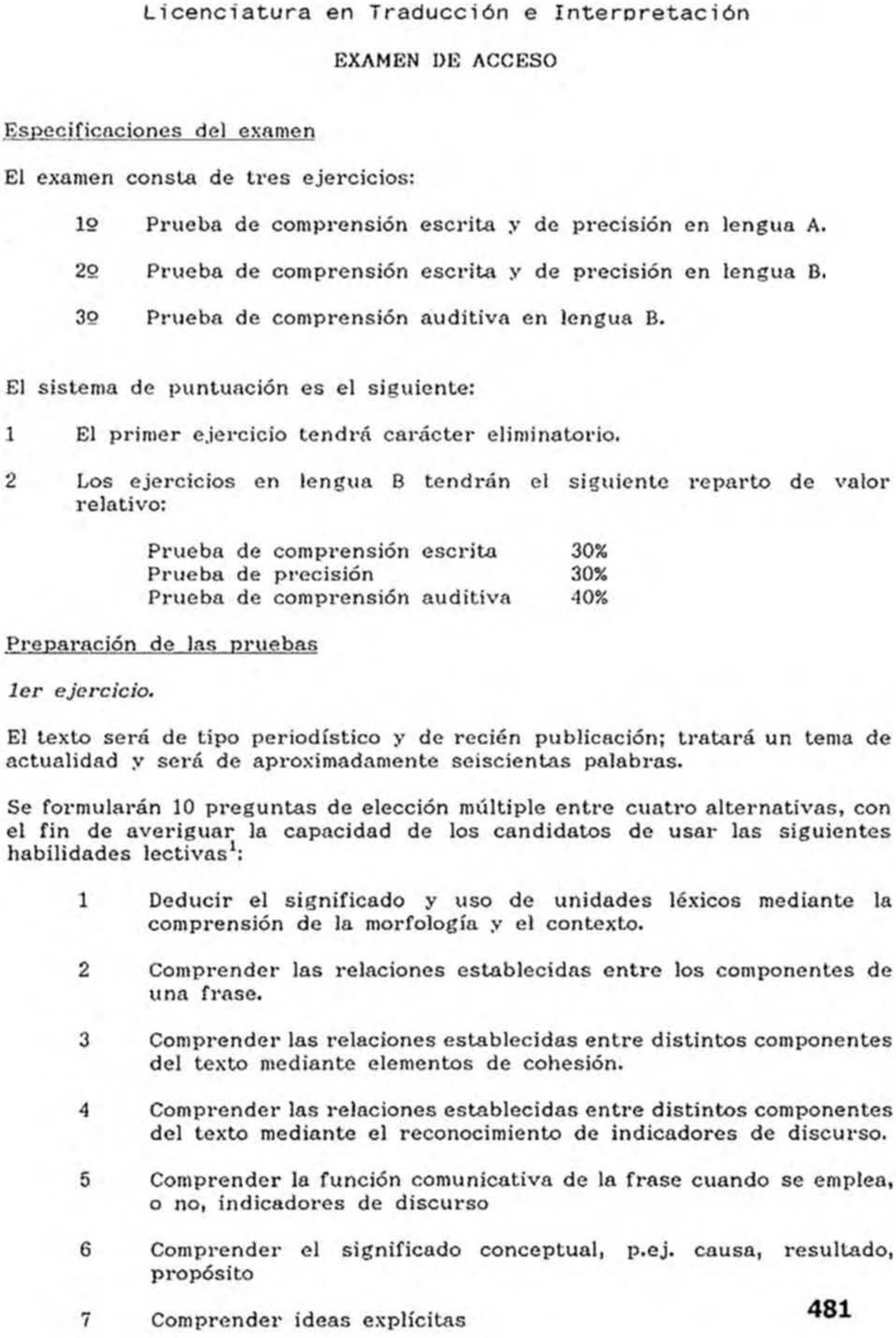
Item n°	DI
19	0.59
20	0.52
14	0.51
11	0.49
13	0.33
6	0.31
5	0.31
8	0.22
9	0.21
4	0.18
16	0.17
2	0.09
1	0.09
7	0.06
10	0.05
15	-0.03
12	-0.06
17	-0.06
3	-0.13
18	-0.17

Rank-order by Facility value

Item nº	FV
17B	0.12
10?	0.12
7?	0.14
5B	0.28
15C	0.31
12E	0.36
13D	0.41
16A	0.46
18E	0.48
11A	0.52
9?	0.56
8V	0.63
4F	0.67
6V	0.71
14E	0.74
2C	0.74
20C	0.75
3D	0.77
19B	0.87
1A	0.91

Test specification

Cohort B



Licenciatura en Traducción e Interpretación

EXAMEN DE ACCESO

Especificaciones del examen

El examen consta de tres ejercicios:

1º Prueba de comprensión escrita y de precisión en lengua A.

2º Prueba de comprensión escrita y de precisión en lengua B.

3º Prueba de comprensión auditiva en lengua B.

El sistema de puntuación es el siguiente:

1 El primer ejercicio tendrá carácter eliminatorio.

2 Los ejercicios en lengua B tendrán el siguiente reparto de valor relativo:

Prueba de comprensión escrita	30%
Prueba de precisión	30%
Prueba de comprensión auditiva	40%

Preparación de las pruebas

1er ejercicio.

El texto será de tipo periodístico y de recién publicación; tratará un tema de actualidad y será de aproximadamente seiscientas palabras.

Se formularán 10 preguntas de elección múltiple entre cuatro alternativas, con el fin de averiguar la capacidad de los candidatos de usar las siguientes habilidades lectivas[1]:

1 Deducir el significado y uso de unidades léxicos mediante la comprensión de la morfología y el contexto.

2 Comprender las relaciones establecidas entre los componentes de una frase.

3 Comprender las relaciones establecidas entre distintos componentes del texto mediante elementos de cohesión.

4 Comprender las relaciones establecidas entre distintos componentes del texto mediante el reconocimiento de indicadores de discurso.

5 Comprender la función comunicativa de la frase cuando se emplea, o no, indicadores de discurso

6 Comprender el significado conceptual, p.ej. causa, resultado, propósito

7 Comprender ideas explícitas

8 Comprender ideas implícitas

9 Separar el contenido esencial de lo no esencial; distinguir entre las ideas principales y los detalles que las ilustren.

A continuación habrá un ejercicio de precisión sobre uno o más párrafos del mismo texto. Se introducirá en el texto unos diez errores de grámatica, ortografia, puntuación y/o de cohesión. El candidato tendrá que indicar la(s) palabra(s) que contienen el error y sugerir correcciones. En total serán 20 ítems, 10 de cada clase.

2º ejercicio.

El texto será de tipo periodístico y de recién publicación; tratará un tema de actualidad, pero no de opinión, y será de aproximadamente cuatrocientas palabras a 12 carácteres por pulgada. También constará de 10 ítems de elecciόm múltiple basados en las mismas habilidades, y de un ejercicio de precisión.

3er ejercicio.

El texto será un monólogo grabado de una fuente pública de difusión - radio o televisión. Tratará un tema de interés general, sin ser ni de opinión, ni sobre un tema semi-especializado. El total de la grabación durará entre dos y cuatro minutos. El acento del presentador será de una variedad de uso común a nivel internacional: p. ej. Inglés británico e inglés americano.

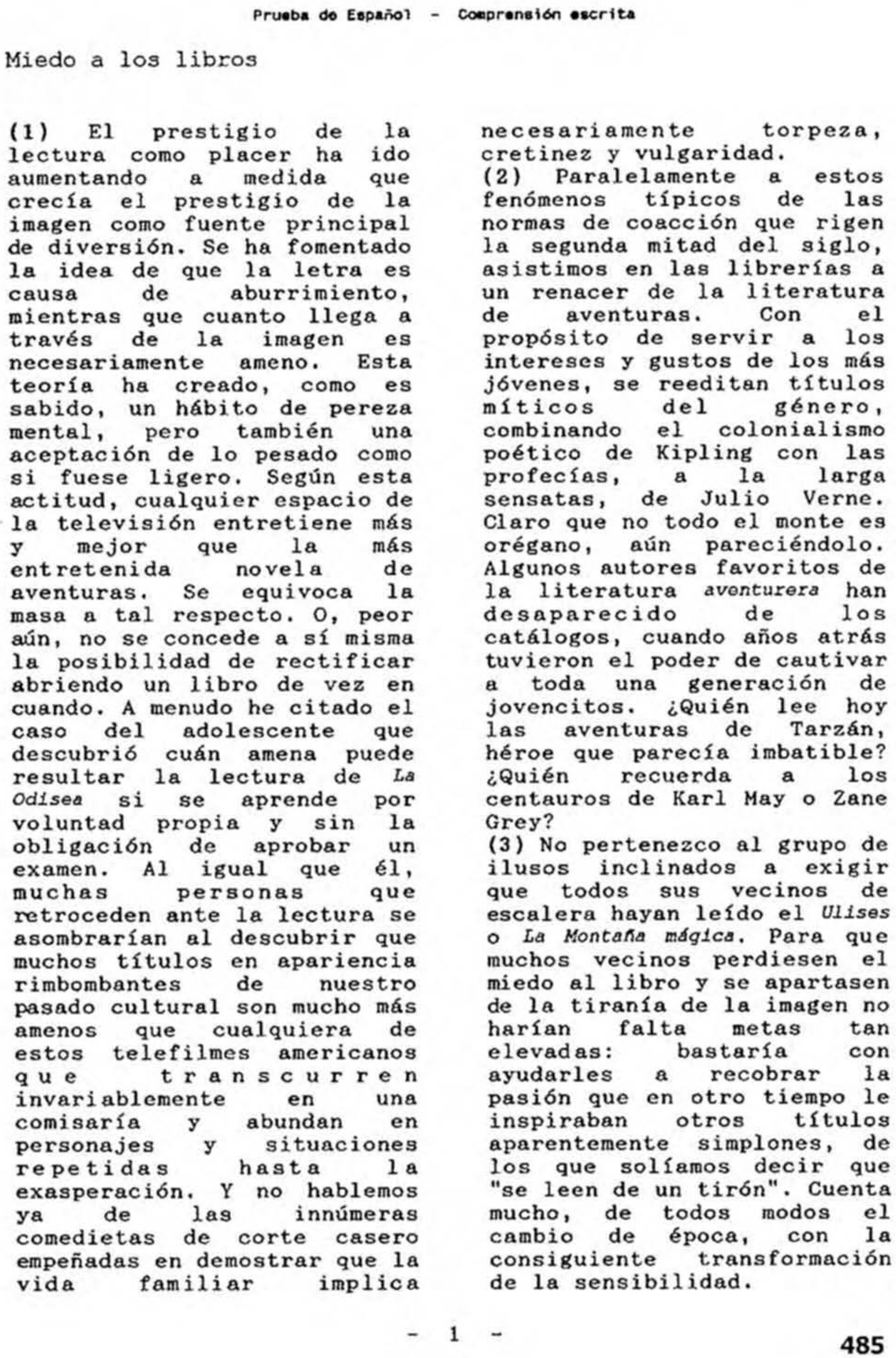
Se formularán las preguntas con el fin de averiguar la capacidad de los candidatos de sintetizar el contenido del texto.

Serán 10 ítems de elección múltiple con tres alternativas, siempre en base a la síntesis de los puntos de contenido más importantes del texto.

Los candidatos podrán escuchar la grabación dos veces y tomar los apuntes que creen pertinentes acerca del contenido del texto, antes de ver las preguntas. A continuación dedicarán un máximo de quince minutos a las respuestas. No volverán a oir la grabación una vez que hayan comenzado a contestar las preguntas.

[1] MUNBY, J.: *Communicative Syllabus Design.*

Sub-test es04



Miedo a los libros

(1) El prestigio de la lectura como placer ha ido aumentando a medida que crecía el prestigio de la imagen como fuente principal de diversión. Se ha fomentado la idea de que la letra es causa de aburrimiento, mientras que cuanto llega a través de la imagen es necesariamente ameno. Esta teoría ha creado, como es sabido, un hábito de pereza mental, pero también una aceptación de lo pesado como si fuese ligero. Según esta actitud, cualquier espacio de la televisión entretiene más y mejor que la más entretenida novela de aventuras. Se equivoca la masa a tal respecto. O, peor aún, no se concede a sí misma la posibilidad de rectificar abriendo un libro de vez en cuando. A menudo he citado el caso del adolescente que descubrió cuán amena puede resultar la lectura de *La Odisea* si se aprende por voluntad propia y sin la obligación de aprobar un examen. Al igual que él, muchas personas que retroceden ante la lectura se asombrarían al descubrir que muchos títulos en apariencia rimbombantes de nuestro pasado cultural son mucho más amenos que cualquiera de estos telefilmes americanos que transcurren invariablemente en una comisaría y abundan en personajes y situaciones repetidas hasta la exasperación. Y no hablemos ya de las innúmeras comedietas de corte casero empeñadas en demostrar que la vida familiar implica necesariamente torpeza, cretinez y vulgaridad.

(2) Paralelamente a estos fenómenos típicos de las normas de coacción que rigen la segunda mitad del siglo, asistimos en las librerías a un renacer de la literatura de aventuras. Con el propósito de servir a los intereses y gustos de los más jóvenes, se reeditan títulos míticos del género, combinando el colonialismo poético de Kipling con las profecías, a la larga sensatas, de Julio Verne. Claro que no todo el monte es orégano, aún pareciéndolo. Algunos autores favoritos de la literatura *aventurera* han desaparecido de los catálogos, cuando años atrás tuvieron el poder de cautivar a toda una generación de jovencitos. ¿Quién lee hoy las aventuras de Tarzán, héroe que parecía imbatible? ¿Quién recuerda a los centauros de Karl May o Zane Grey?

(3) No pertenezco al grupo de ilusos inclinados a exigir que todos sus vecinos de escalera hayan leído el *Ulises* o *La Montaña mágica*. Para que muchos vecinos perdiesen el miedo al libro y se apartasen de la tiranía de la imagen no harían falta metas tan elevadas: bastaría con ayudarles a recobrar la pasión que en otro tiempo le inspiraban otros títulos aparentemente simplones, de los que solíamos decir que "se leen de un tirón". Cuenta mucho, de todos modos el cambio de época, con la consiguiente transformación de la sensibilidad.

Elige la alternativa más apropiada y subraya la letra que corresponda en la hoja de respuestas.

1. Según el texto, la idea de que la lectura provoca aburrimiento, (párrafo 1) ¿qué hábito ha creado entre la gente?

   A Potenciar la cultura de la imagen.
   B Dificultad de concentrarse.
   C Acostarse más temprano.
   D Ver más la TV.

2. "Cualquier espacio de la TV entretiene más y mejor que la más entretenida novela de aventuras ..." (párrafo 1) ¿qué figura retórica parece que emplea el autor en esta frase?

   A Anáfora.
   B Paradoja.
   C Anacoluto.
   D Redundancia.

3. "... muchos títulos en apariencia <u>rimbombantes</u> ..." (párrafo 1) ¿qué adjetivo estaría más próximo a la palabra subrayada?

   A Sonoro.
   B Rumboso.
   C Glorioso.
   D Llamativo.

4. "Comedietas de corte casero ..." (párrafo 1). Con esta expresión el autor se refiere ...

   A al uso despectivo del término comedia.
   B a la Comedia de costumbres.
   C a la Comedia de carácter.
   D a la Comedia familiar.

5. "... se reeditan <u>títulos míticos</u> del género ..." (párrafo 2). Con esta frase el autor se refiere a los libros ...

   A de profundo sentido religioso.
   B más conocidas.
   C de dioses griegos.
   D más consultadas por los especialistas.

6. "... las profecías, a la larga sensatas, ..." (párrafo 2). La expresión subrayada se podría sustituir por ...

A cuando ha pasado un tiempo.
B a su debido tiempo.
C en su momento.
D siempre.

7. "... tuvieron el poder de cautivar a toda una generación ..." (párrafo 2). ¿Cuál sería el verbo más exactamente contrario de cautivar?

A Desencantar.
B Aborrecer.
C Rechazar.
D Repeler.

8. "Quién recuerda a los centauros de Karl May ..." (párrafo 2). Con qué palabra o palabras podríamos sustituir el término "centauros" según el sentido del texto.

A Vaqueros (cow-boys).
B Jinetes audaces.
C Policía montada.
D Caballeros.

9. El haber leído *Ulises* o *La Montaña mágica* (3) supone ...

A ser un experto en Literatura griega.
B ser aficionados a los cuentos clásicos.
C estar al día en la Literatura contemporánea.
D gustar los libros de aventuras.

10. " ... la tiranía de la imagen ..." (párrafo 3) se refiere a la ...

A carencia de otros intereses visuales.
B ausencia de otros medios.
C atracción absoluta hacia el cine y la TV.
D influencia de los medios audiovisuales.

A continuación, presentamos el último párrafo del texto; en él se han introducido 10 errores de tipo lingüístico (acentuación, ortografía, morfosintaxis y puntuación). Marcar el punto donde se encuentre una incorrección y proponed la solución correcta en la línea de al lado.

Comprendo que al enfrentar ejemplos de la cultura llamada "superior" contra los de la supcultura, me adentro en una selva tan intricada como la de la Comedia. En cualquier caso, conviene disponer de la suficiente humildad para no despotricar contra las novelas que, segun el tópico, "se leen de un tirón". En realidad, restituyan al lector perezoso sus derechos sobre la imaginación, hoy aletardada bajo los sobornos constantes de la caja tonta. Este esfuerzo es el primero que cabría esigir. Sigo recordando el ejemplo de mi amiguito el adolescente homeréfilo, salió de *La Odisea* tan encantado que se puso de inmediato a buscar experiencias parecidas a otros libros. Mientras se perdió quince o veinte concursos de varias cadenas. Y si hizo esto un pobre adolescente despistado, ¿han de ser de menos mis vecinos adultos?.

Prueba de Español - Comprensión escrita

Miedo a los libros

(1) El prestigio de la lectura como placer ha ido aumentando a medida que crecía el prestigio de la imagen como fuente principal de diversión. Se ha fomentado la idea de que la letra es causa de aburrimiento, mientras que cuanto llega a través de la imagen es necesariamente ameno. Esta teoría ha creado, como es sabido, un hábito de pereza mental, pero también una aceptación de lo pesado como si fuese ligero. Según esta actitud, cualquier espacio de la televisión entretiene más y mejor que la más entretenida novela de aventuras. Se equivoca la masa a tal respecto. O, peor aún, no se concede a sí misma la posibilidad de rectificar abriendo un libro de vez en cuando. A menudo he citado el caso del adolescente que descubrió cuán amena puede resultar la lectura de *La Odisea* si se aprende por voluntad propia y sin la obligación de aprobar un examen. Al igual que él, muchas personas que retroceden ante la lectura se asombrarían al descubrir que muchos títulos en apariencia rimbombantes de nuestro pasado cultural son mucho más amenos que cualquiera de estos telefilmes americanos que transcurren invariablemente en una comisaría y abundan en personajes y situaciones repetidas hasta la exasperación. Y no hablemos ya de las innúmeras comedietas de corte casero empeñadas en demostrar que la vida familiar implica necesariamente torpeza, cretinez y vulgaridad.

(2) Paralelamente a estos fenómenos típicos de las normas de coacción que rigen la segunda mitad del siglo, asistimos en las librerías a un renacer de la literatura de aventuras. Con el propósito de servir a los intereses y gustos de los más jóvenes, se reeditan títulos míticos del género, combinando el colonialismo poético de Kipling con las profecías, a la larga sensatas, de Julio Verne. Claro que no todo el monte es orégano, aún pareciéndolo. Algunos autores favoritos de la literatura *aventurera* han desaparecido de los catálogos, cuando años atrás tuvieron el poder de cautivar a toda una generación de jovencitos. ¿Quién lee hoy las aventuras de Tarzán, héroe que parecía imbatible? ¿Quién recuerda a los centauros de Karl May o Zane Grey?

(3) No pertenezco al grupo de ilusos inclinados a exigir que todos sus vecinos de escalera hayan leído el *Ulises* o *La Montaña mágica*. Para que muchos vecinos perdiesen el miedo al libro y se apartasen de la tiranía de la imagen no harían falta metas tan elevadas: bastaría con ayudarles a recobrar la pasión que en otro tiempo le inspiraban otros títulos aparentemente simplones, de los que solíamos decir que "se leen de un tirón". Cuenta mucho, de todos modos el cambio de época, con la consiguiente transformación de la sensibilidad.

Prueba de Español - Comprensión escrita

1er APELLIDO:

2º APELLIDO:

NOMBRE:

DNI:

[TOTAL:]

Subraya la letra o símbolo que corresponda.

1.	A	B	C	D
2.	A	B	C	D
3.	A	B	C	D
4.	A	B	C	D
5.	A	B	C	D
6.	A	B	C	D
7.	A	B	C	D
8.	A	B	C	D
9.	A	B	C	D
10.	A	B	C	D

Prueba de Español - Comprensión escrita

Elige la alternativa más apropiada y subraya la letra que corresponda en la hoja de respuestas.

1. Según el texto, la idea de que la lectura provoca aburrimiento, (párrafo 1) ¿qué hábito ha creado entre la gente?

   A Potenciar la cultura de la imagen.
   B Dificultad de concentrarse.
   C Acostarse más temprano.
   D Ver más la TV.

2. "Cualquier espacio de la TV entretiene más y mejor que la más entretenida novela de aventuras ..." (párrafo 1) ¿qué figura retórica parece que emplea el autor en esta frase?

   A Anáfora.
   B Paradoja.
   C Anacoluto.
   D Redundancia.

3. "... muchos títulos en apariencia rimbombantes ..." (párrafo 1) ¿qué adjetivo estaría más próximo a la palabra subrayada?

   A Sonoro.
   B Rumboso.
   C Glorioso.
   D Llamativo.

4. "Comedietas de corte casero ..." (párrafo 1). Con esta expresión el autor se refiere ...

   A al uso despectivo del término comedia.
   B a la Comedia de costumbres.
   C a la Comedia de carácter.
   D a la Comedia familiar.

5. "... se reeditan títulos míticos del género ..." (párrafo 2). Con esta frase el autor se refiere a los libros ...

   A de profundo sentido religioso.
   B más conocidos.
   C de dioses griegos.
   D más consultados por los especialistas.

6. "... las profecías, a la larga sensatas, ..." (párrafo 2). La expresión subrayada se podría sustituir por ...

A cuando ha pasado un tiempo.
B a su debido tiempo.
C en su momento.
D siempre.

7. "... tuvieron el poder de cautivar a toda una generación ..." (párrafo 2). ¿Cuál sería el verbo más exactamente contrario de cautivar?

A Desencantar.
B Aborrecer.
C Rechazar.
D Repeler.

8. "Quién recuerda a los centauros de Karl May ..." (párrafo 2). Con qué palabra o palabras podríamos sustituir el término "centauros" según el sentido del texto.

A Vaqueros (cow-boys).
B Jinetes audaces.
C Policía montada.
D Caballeros.

9. El haber leído *Ulises* o *La Montaña mágica* **(3)** supone ...

A ser un experto en Literatura griega.
B ser aficionados a los cuentos clásicos.
C estar al día en la Literatura contemporánea.
D gustar los libros de aventuras.

10. " ... la tiranía de la imagen ..." (párrafo 3) se refiere a la ...

A carencia de otros intereses visuales.
B ausencia de otros medios.
C atracción absoluta hacia el cine y la TV.
D influencia de los medios audiovisuales.

Prueba de Español - Comprensión escrita

A continuación, presentamos el último párrafo del texto; en él se han introducido 10 errores de tipo lingüístico (acentuación, ortografía, morfosintaxis y puntuación). Marcar el punto donde se encuentre una incorrección y proponed la solución correcta en la línea de al lado.

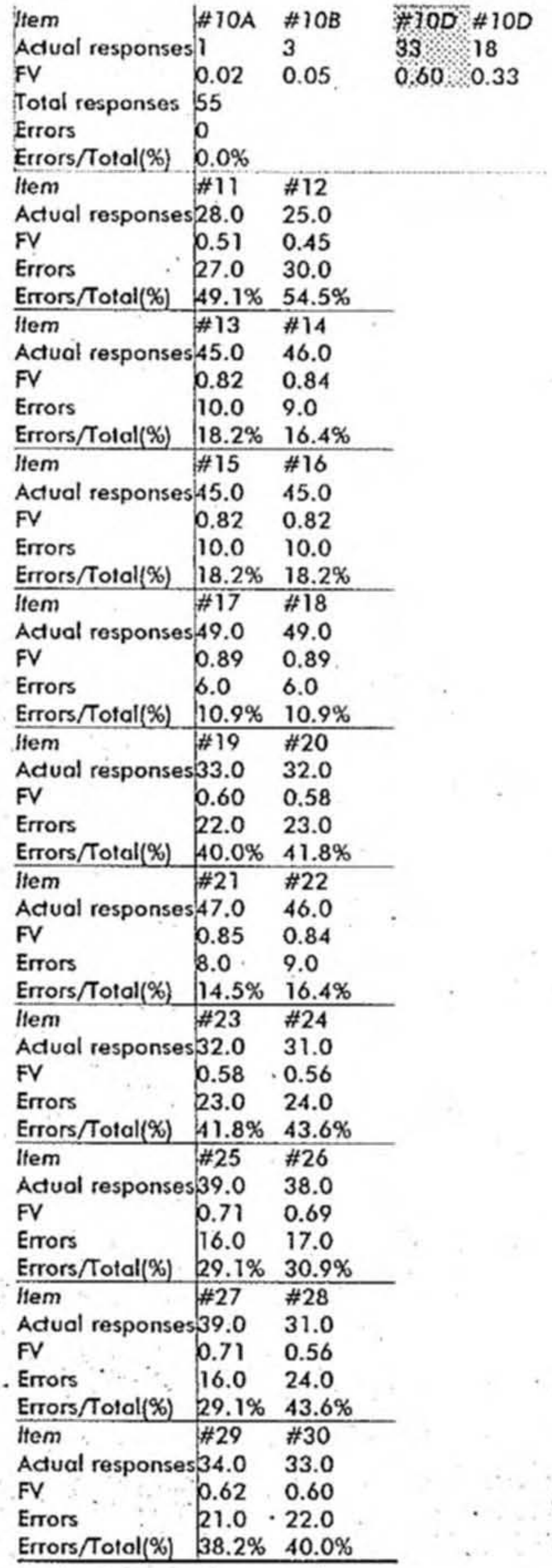
Comprendo ~~que al~~ enfrentar ejemplos de la cultura llamada "superior" contra los de la ~~supecultura~~, me adentro en una selva tan intricada como la de la Comedia. En cualquier caso, conviene disponer de la suficiente humildad para no despotricar contra las novelas que, ~~segun~~ el tópico, "se leen de un tirón". En realidad, ~~restituyan~~ al lector perezoso sus derechos sobre la imaginación, hoy ~~aletardada~~ bajo los sobornos constantes de la caja tonta. Este esfuerzo es el primero que cabría ~~esigir~~. Sigo recordando el ejemplo de mi amiguito el adolescente ~~homerófilo;~~ salió de *La Odisea* tan encantado que se puso de inmediato a buscar experiencias parecidas ~~a~~ otros libros. ~~Mientras~~ se perdió quince o veinte concursos de varias cadenas. Y si hizo esto un pobre adolescente despistado, ¿han de ~~ser de menos~~ mis vecinos adultos?.

Comprendo que, al enfrentar ejemplos de la cultura llamada "superior" contra los de la subcultura, me adentro en una selva tan intricada como la de la Comedia. En cualquier caso, conviene disponer de la suficiente humildad para no despotricar contra las novelas que, según el tópico, "se leen de un tirón". En realidad, restituyen al lector perezoso sus derechos sobre la imaginación, hoy aletargada bajo los sobornos constantes de la caja tonta. Este esfuerzo es el primero que cabría exigir. Sigo recordando el ejemplo de mi amiguito el adolescente homerófilo; salió de *La Odisea* tan encantado que se puso de inmediato a buscar experiencias parecidas en otros libros. Mientras, se perdió quince o veinte concursos de varias cadenas. Y si hizo esto un pobre adolescente despistado, ¿han de ser menos mis vecinos adultos?.

Table1 Test es04 Printout of Facility values (FVs)

Item	*#1A*	*#1B*	*#1C*	*#1D*
Actual responses	29	6	0	20
FV	0.53	0.11	0.00	0.36
Total responses	55			
Errors	0			
Errors/Total(%)	0.0%			
Item	*#2A*	*#2B*	*#2C*	*#2D*
Actual responses	5	14	6	30
FV	0.09	0.25	0.11	0.55
Total responses	55			
Errors	0			
Errors/Total(%)	0.0%			
Item	*#3A*	*#3B*	*#3C*	*#3D*
Actual responses	15	9	5	26
FV	0.27	0.16	0.09	0.47
Total responses	55			
Errors	0			
Errors/Total(%)	0.0%			
Item	*#4A*	*#4B*	*#4C*	*#4D*
Actual responses	28	5	0	22
FV	0.51	0.09	0.00	0.40
Total responses	55			
Errors	0			
Errors/Total(%)	0.0%			
Item	*#5A*	*#5B*	*#5C*	*#5D*
Actual responses	0	51	4	0
FV	0.00	0.93	0.07	0.00
Total responses	55			
Errors	-3			
Errors/Total(%)	-8.1%			
Item	*#6A*	*#6B*	*#6C*	*#6D*
Actual responses	32	14	3	7
FV	0.58	0.25	0.05	0.13
Total responses	56			
Errors	-1			
Errors/Total(%)	-1.8%			
Item	*#7A*	*#7B*	*#7C*	*#7D*
Actual responses	28	9	3	14
FV	0.51	0.16	0.05	0.25
Total responses	54			
Errors	1			
Errors/Total(%)	1.8%			
Item	*#8A*	*#8B*	*#8C*	*#8D*
Actual responses	8	29	2	16
FV	0.15	0.53	0.04	0.29
Total responses	55			
Errors	0			
Errors/Total(%)	0.0%			
Item	*#9A*	*#9B*	*#9C*	*#9D*
Actual responses	9	15	13	18
FV	0.16	0.27	0.24	0.33
Total responses	55			
Errors	0			
Errors/Total(%)	0.0%			

2



Item	*#10A*	*#10B*	*#10D*	*#10D*
Actual responses	1	3	33	18
FV	0.02	0.05	0.60	0.33
Total responses	55			
Errors	0			
Errors/Total(%)	0.0%			
Item	#11	#12		
Actual responses	28.0	25.0		
FV	0.51	0.45		
Errors	27.0	30.0		
Errors/Total(%)	49.1%	54.5%		
Item	#13	#14		
Actual responses	45.0	46.0		
FV	0.82	0.84		
Errors	10.0	9.0		
Errors/Total(%)	18.2%	16.4%		
Item	#15	#16		
Actual responses	45.0	45.0		
FV	0.82	0.82		
Errors	10.0	10.0		
Errors/Total(%)	18.2%	18.2%		
Item	#17	#18		
Actual responses	49.0	49.0		
FV	0.89	0.89		
Errors	6.0	6.0		
Errors/Total(%)	10.9%	10.9%		
Item	#19	#20		
Actual responses	33.0	32.0		
FV	0.60	0.58		
Errors	22.0	23.0		
Errors/Total(%)	40.0%	41.8%		
Item	#21	#22		
Actual responses	47.0	46.0		
FV	0.85	0.84		
Errors	8.0	9.0		
Errors/Total(%)	14.5%	16.4%		
Item	#23	#24		
Actual responses	32.0	31.0		
FV	0.58	0.56		
Errors	23.0	24.0		
Errors/Total(%)	41.8%	43.6%		
Item	#25	#26		
Actual responses	39.0	38.0		
FV	0.71	0.69		
Errors	16.0	17.0		
Errors/Total(%)	29.1%	30.9%		
Item	#27	#28		
Actual responses	39.0	31.0		
FV	0.71	0.56		
Errors	16.0	24.0		
Errors/Total(%)	29.1%	43.6%		
Item	#29	#30		
Actual responses	34.0	33.0		
FV	0.62	0.60		
Errors	21.0	22.0		
Errors/Total(%)	38.2%	40.0%		

Table 1 Test es04 Printout of Discrimination indices (DIs)

Item	*#1A*	*#2B*	*#3D*	*#4A*	*#5B*	*#6A*	*#7D*	*#8A*	*#9C*	*#10C*
H	10	3	7	8	13	12	4	3	6	14
L	8	3	6	6	14	5	3	1	1	6
n	15	15	15	15	15	15	15	15	15	15
DI	0.13	0.00	0.07	0.13	-0.07	0.47	0.07	0.13	0.33	0.53
Item	*#1B*	*#2D*								
H	1	9								
L	2	9								
n	15	15								
DI	-0.07	0.00								
Item	*#11*	*#12*	*#13*	*#14*	*#15*	*#16*	*#17*	*#18*	*#19*	*#20*
H	8	8	13	8	14	14	15	15	13	13
L	6	5	12	13	11	11	10	10	6	5
n	15	15	15	15	15	15	15	15	15	15
DI	0.13	0.20	0.07	-0.33	0.20	0.20	0.33	0.33	0.47	0.53
Item	*#21*	*#22*	*#23*	*#24*	*#25*	*#26*	*#27*	*#28*	*#29*	*#30*
H	15	15	10	10	13	13	12	9	14	14
L	11	11	5	5	7	6	7	6	6	6
n	15	15	15	15	15	15	15	15	15	15
DI	0.27	0.27	0.33	0.33	0.40	0.47	0.33	0.20	0.53	0.53

H = Correct responses by candidates in higher 27.5% of cohort

L = Correct responses by candidates in lower 27.5% of cohort

n = 27.5% of cohort

INFORME

Traducción e Interpretación

Prueba específica de aptitud

Universidad Alfonso X El Sabio, Junio-Julio 1994.

A Introducción.

Este informe consta de una serie de datos acerca de cada ejercicio, las conclusiones que se pueden derivar de ellos, comentarios acerca de algunos candidatos individuales y sugerencias sobre los grupos de Lengua B que se podrían formar una vez empezado el curso académico 1994-95. Al final se presentan de forma gráfica los datos más importantes sobre los resultados.

Los resultados de las pruebas en Lengua A (Español) y Lengua B (Inglés) se basan en un número de candidatos suficiente para poder inferir conclusiones; las pruebas en Lengua B (Francés) fueron realizadas por un grupo demasiado pequeño para un análisis estadístico; no se presentaron candidatos para Lengua B (Alemán).

B Lengua A (Español): Comprensión escrita y precisión.

La prueba en Lengua A (Español) se compone de dos ejercicios que puntúan independientemente y que luego hacen media para formar la nota final (%).

Ambos ejercicios resultaron difíciles para este grupo de candidatos y, aunque no se puede descartar el hecho de que la prueba de aptitud haya sido algo inesperado para ellos, se debe suponer que las notas de todos estos candidatos en la prueba de Lengua Española de Selectividad - que este año se asemeja a la prueba que hemos realizado -hayan sido relativamente bajos.

En Lengua (A) no se ha fijado ninguna nota de "Aprobado".

El coeficiente de correlación entre los dos ejercicios demuestra que estos han determinado niveles de competencia en habilidades completamente distintas.

C Lengua B (Inglés): Comprensión escrita, precisión, y comprensión oral.

Los dos primeros ejercicios - comprensión escrita y precisión - han resultado bastante difíciles para este grupo de candidatos; el tercer ejercicio ha producido resultados satisfactorios.

En el caso de cada componente se ha sugerido una nota de aprobado como medida orientativa. Esta nota se calcula al restar la desviación típica (DT) de la puntuación o nota media del grupo de candidatos. (Así se fijaron los niveles de aprobado en la Universidad de Granada para la primera promoción de la Licenciatura en Traducción e Interpretación para el curso 1993-94.)

Los coeficientes de correlación entre los tres ejercicios demuestran que estos han determinado niveles de competencia en habilidades completamente distintas, por lo cual en el caso de cada candidato puede resultar interesante consultar las distintas puntuaciones adquiridas.

D Lengua B (Francés) Comprensión escrita, precisión y comprensión oral.

El cómputo de las puntuaciones y de la nota final se ha realizado de la misma manera que en el caso de Lengua B (Inglés).

Los siete candidatos parecen haber encontrado el examen más bien fácil y no existen grandes diferencias entre las notas que reciben.

E Candidatos individuales:

E.1 Lengua B (Inglés)

BALLESTEROS RABANO, ELENA
CASTELLO MARTINEZ, MIGUEL ÁNGEL
MANZANO CUEVAS, PASCUALA ISABEL
RODRÍGUEZ CORRAL, LUCIA
VIDAL GROHALL, ANA

Todos los arriba mencionados han sacado notas bajas en Lengua A; sus notas en Lengua B son bajas, tampoco destacan en ninguno de los componentes de la prueba. *Se sugiere que antes de admitir a estos candidatos se contrasten las notas que han conseguido en estas pruebas con las de la Selectividad y con los resultados de los tests psicotécnicos.*

QUINN, PETER

Este candidato - irlandès de nacionalidad - ha sacado un nota muy baja en Lengua A (Español) y alta en Lengua B (Inglés). No obstante, al estudiar los tres componentes de la nota de Lengua B se ve que su fuerte es la Comprensión Oral, lo que es lógico. *Se sugiere que antes de admitir a este candidato se contraste las notas que ha conseguido en éstas pruebas con las de la Selectividad y con los resultados de los tests psicotécnicos.*

E.2 Lengua B (Francés)

CAMPOS PITA, ISABEL SUSANA

Esta candidata no presentó respuesta alguna al primer ejercicio de Lengua B (Francés). A pesar de que es probable que casi no exista correlación entre los elementos de la prueba, se podría suponer que tiene un nivel suficientemente alto como para entrar en la promoción. *Se sugiere admitirla sin más.*

F Grupos de Lengua A (Español).

Aunque, de entrada, no sería necesario tener en cuenta grupos de Lengua A por niveles de aptitud, se debería de pensar en la necesidad de proporcionarles a algunos de los candidatos trabajos complementarios. Éstos son:

CASTELLO MARTINEZ, MIGUEL ÁNGEL
CLAVERO AMOR, EMMA
FRANCISCO GÓMEZ-ESCOLAR, PATRICIA DE
GONZÁLEZ AGUADO, ARANZAZU
QUINN, PETER
TORRES MARTINEZ, RAQUEL

G Grupos de Lengua B (Inglés).

Si se decidiera admitir a todos los candidatos la división inicial, en tres grupos de niveles homogéneos de aptitud, podría ser:

Nivel A

BALLESTEROS RABANO, ELENA
CASTELLO MARTINEZ, MIGUEL ÁNGEL
MANZANO CUEVAS, PASCUALA ISABEL
SALVADOR PRADA, EVA CANDELAS
RODRÍGUEZ CORRAL, LUCIA
VIDAL GROHALL, ANA

Nivel B

COBO PÉREZ, IGNACIO
ESPINOSA WILHELMI, JUAN MANUEL
FRANCISCO GÓMEZ-ESCOLAR, PATRICIA DE
GARCÍA RODRÍGUEZ, GISELA
GARCÍA RUBIO, ANA
GONZÁLEZ AGUADO, ARANZAZU
GONZÁLEZ DÍAZ, VICTORINA
GONZÁLEZ DIOS, ANA
GULLON PÉREZ, MARÍA JESÚS
LÓPEZ BERROTARAN, SILVIA
LOZANO DE LA CRUZ, AGUSTÍN
PUENTE RINCÓN, ANA BELÉN DE
VEGA LÓPEZ, SILVIA DE LA

VILLAMAYOR HIERRO, BERTA M

Nivel C

ARLT GHIO, CHRISTIAN
CASTRO ARELLANO, ISABEL M DE
CENDON GARCÍA DE LEANIZ, MARGOT
CLAVERO AMOR, EMMA
DÍAZ GARCÍA-VERDUGO, LAURA
GALLEGO GARCÍA, JESÚS
GARCÍA BLANCO, ÁNGELES
GORBEA SALVADOR, ANAHI
GRACIA VICENTE, MARTA
JIMÉNEZ AGUEDA, MARTA
LANDALUCE ABURTO, NORA
LÒPEZ FERNÁNDEZ, M PAZ
MARCOS GARRIDO, GEMA
MARTINEZ RUIZ, JUAN RAMÓN
MELCON MARTINEZ, PATRICIA
MENDOZA MARTÍN, BELÉN
MERLINO SÁNCHEZ-ELVIRA, MIRIAM
MIGUEZ PÉREZ, LUISA
PONTE CATENA, CRISTINA
PRIETO HERRERO, M DEL MAR
PUERTAS CERDEÑO, M DEL MAR
QUINN, PETER
RIVERA CAVANILLES, JULIO
ROJO CABRERA, ADRIANA
SÁNCHEZ VAZQUEZ, NATALIA
SANZ-BUSTILLO BUTRAGUEÑO, ESTEBAN
TORRES MARTINEZ, RAQUEL
VERDU CANO, CATALINA

Estos grupos deberían de reajustarse a lo largo o al final del primer trimestre.

H Grupos de Lengua B (Francés).

Se supone que habrá un solo grupo en Francés.

I Resumen de datos.

I.1 Lengua A (Español): Comprensión escrita y precisión.

N□ candidatos	55		
	Comprensión Escrita (/10)	Precisión (/20)	Total (%)
Nota Max	7	9	58
Nota Min	1	3	20
Rango	6	6	38
Media	4.09	6.7	37
Moda	5	8	40
Mediana	4	7	39
Desviación típica	1.3	1.5	7.8
Nota aprobado (Media menos 1 DT)	N/A	N/A	30
Coeficiente de correlación (CE v. Pr)	0.104		

I.2 Lengua B (Inglés): Comprensión escrita, precisión, y comprensión oral.

Nº candidatos	48			
	Comp. Escrita (/10)	Precisión (/20)	Comp. Oral (/10)	Total (8)
Nota Max	7	11	10	74
Nota Min	0	0	1	13
Rango	7	11	9	61
Media	3.83	4.5	5.83	44
Moda	5	5	6	50
Mediana	4	5	6	46
Desviación típica	1.8	2.56	1.67	11
Nota aprobado (Media menos 1 DT)	2.03	1.94	4.16	33
Coef. de cor-relación	**CE v. Pr** 0.054	**Pr v. CO** 0.302	**CE v. CO** 0.013	**N/A**

Sub-test in03

A skeleton in the cupboard

Birna Helgadottir on the literary triumph of an icelander who is also a New York yuppie prince

Olafur Olafsson
Absolution *Orion, £12.99*

(1) IF OLAFUR OLAFSSON were a character in one of his own novels, critics would no doubt think him a far-fetched fantasy. The son of one of Iceland's best-known writers, Olafur Johann Sigurdsson, the young Olafsson joined Sony, rising to become, at the height of New York yuppie fever, the president of Sony Electronic Publishing - one of the youngest company presidents in US business history. At the same time, he published his first novel to rave reviews back in Iceland.

(2) Now, at 31, both Olafsson the executive and Olafsson the novelist are flourishing to an extraordinary degree. Electronic Publishing, riding the crest of the video game boom, is the fastest-growing branch of the Sony empire. Meanwhile its president, already the best-selling author in Iceland's history has had his first novel in English, *Absolution,* hailed by critics in the US and Britain as a worthy successor to Dostoevsky's *Crime and Punishment.*

(3) Olafsson has chosen as the subject of his first "international" work how embittered old age follows a wasted life. *Absolution* is a study of a man whose life has been poisoned by a "little crime" of treachery committed in youth. The novel takes the form of the memoirs of Peter Peterson, an Icelandic expatriate living the life of a recluse in New York's Park Avenue. He has amassed a fortune through a lifetime of dubious business practices, is cruel to his estranged family and has no faith in God or humanity. His only interests are old movies, wine and the charms of his housekeeper.

(4) We are gradually introduced to the old misanthrope's younger self, an idealistic youth, then known by his proper Icelandic name Petur Petursson. Young Petur leaves his native Reykjavik to follow Gudrun, the girl he worships, to Copenhagen, just before the Nazi occupation. It is here, in a blind fog of jealousy and shattered illusion, that he carries out the deed that transforms him into the bitter, amoral Peterson. The narrative, which zips back and forth between past and present, pre-war Scandinavia and modern Manhattan, is sandwiched between the musings of its "editor", a sensitive young Icelander who has discovered Peterson's manuscript after his death.

(5) In Iceland the novel has sold 13,000 copies, one for every 20 inhabitants. Olafsson's success there is in many ways due to the fact that he embodies the curious paradox of the Icelandic psyche.

(6) On the one hand, it is a materialistic nation with one of the highest living standards in the world. In recent years the traditional devotion to the Protestant work ethic has often combined with growing consumerism to produce the worst kind of yuppie excesses.

(7) At the same time, Iceland is probably the world's most literate and literary nation. This is the country that produced not just the medieval sagas - Europe's first vernacular prose - but other more esoteric literary efforts. The world's first Basque dictionary was written by a farmer in North-West Iceland - he had learned the language from Basque fishermen. The only epic poem ever written in honour of the medieval Balkan hero Skanderbeg was composed by an

otherwise unknown Icelandic clergyman in 1861.

(8) Literacy rates have been nearly 100 per cent since at least the 18th century. With 400-500 published here every year, Iceland produces more books per head of population than any other country. One in ten Icelanders will be published in his or her lifetime - writing a book is seen almost a form of national service.

(9) There is a poetic, mystical side to the Icelanders - most claim to believe in God and most believe in elves and ghosts, Poetic ability is considered the hallmark of a noble personality, and artists and writers have a special dispensation, in a nation where hard work is considered beneficial for its own sake, to lead a ramshackle, layabout lifestyle. As the Icelandic Nobel prizewinner Laxness put it, "since time immemorial, the Icelandic nation has had to battle with men who call themselves poets and refuse to work for a living."

(10) But Olafsson is a poet who works hard for his living. He has also fulfilled another important Icelandic dream - to become famous abroad: he has been featured in *Fortune* magazine as a high-flying company president and the *New York Times Book Review* as a promising novelist.

English - Reading Comprehension

Choose the most appropriate of the four alternatives given, and underline the corresponding letter on the answer sheet.

1. The growth of the electronic publishing industry ...

   A continues to amaze the business world.
   B has failed to live up to expectations.
   C appears to have reached a plateau.
   D has past its peak.

2. Critics have compared Olafsson's *Absolution* favourably with ...

   A his previous novels.
   B Dostoevsky's *Crime and Punishment*.
   C the epic poem written in honour of Skanderbeg.
   D the works of the Icelandic Nobel prizewinner Laxness.

3. When the writer uses the words *"little crime"* between inverted commas she does so because she ...

   A is using a recognised literary term.
   B wants to produce an ironic effect.
   C intends it as reported speech.
   D is quoting from the text.

4. *Absolution* ...

   A deals with an Icelandic theme.
   B was first published in Icelandic.
   C is Olafsson's first novel in English.
   D made Olafsson the best-selling author in Iceland's history.

5. In the phrase "... to his estranged family ..." (Paragraph 3), *estranged* means the same as ...

   A unfriendly.
   B foreign.
   C faraway.
   D curious.

English - Reading Comprehension

6. In *Absolution*, the narrative ...

   A displays characteristics typical of Icelandic literature.
   B is a chronological account of the protagonist's life.
   C has been written by an "editor".
   D changes perspective frequently.

7. "... the curious paradox of the Icelandic psyche ... (Paragraph 5) is that as a nation Icelanders ...

   A enjoy high living standards but act like yuppies.
   B are materialistic and yet highly literate and literary.
   C are devout Protestants who have an affinity for the Basques.
   D publish huge numbers of books each year though they read few.

8. One out of ten Icelanders ...

   A has bought a copy of *Absolution* but has not read it.
   B has published a book.
   C has read *Absolution*.
   D is a poet.

9. In Iceland people who think themselves artists or writers ...

   A lead bohemian lives.
   B are social outcasts.
   C refuse to work for a living.
   D believe in elves and ghosts.

10. Olafsson has fulfilled an important Icelandic dream because he ...

   A embodies the traditional image of the writer.
   B does not believe in elves and ghosts.
   C has become famous abroad.
   D works hard.

English - Reading Comprehension

The text continues here, but 10 errors - of grammar (2), cohesion (1), spelling (2), punctuation (3), and vocabulary (2) - have been introduced. Read the text carefully and underline the errors you find. On the line provided in the second column write the correct form. One answer has been given for you as an example.

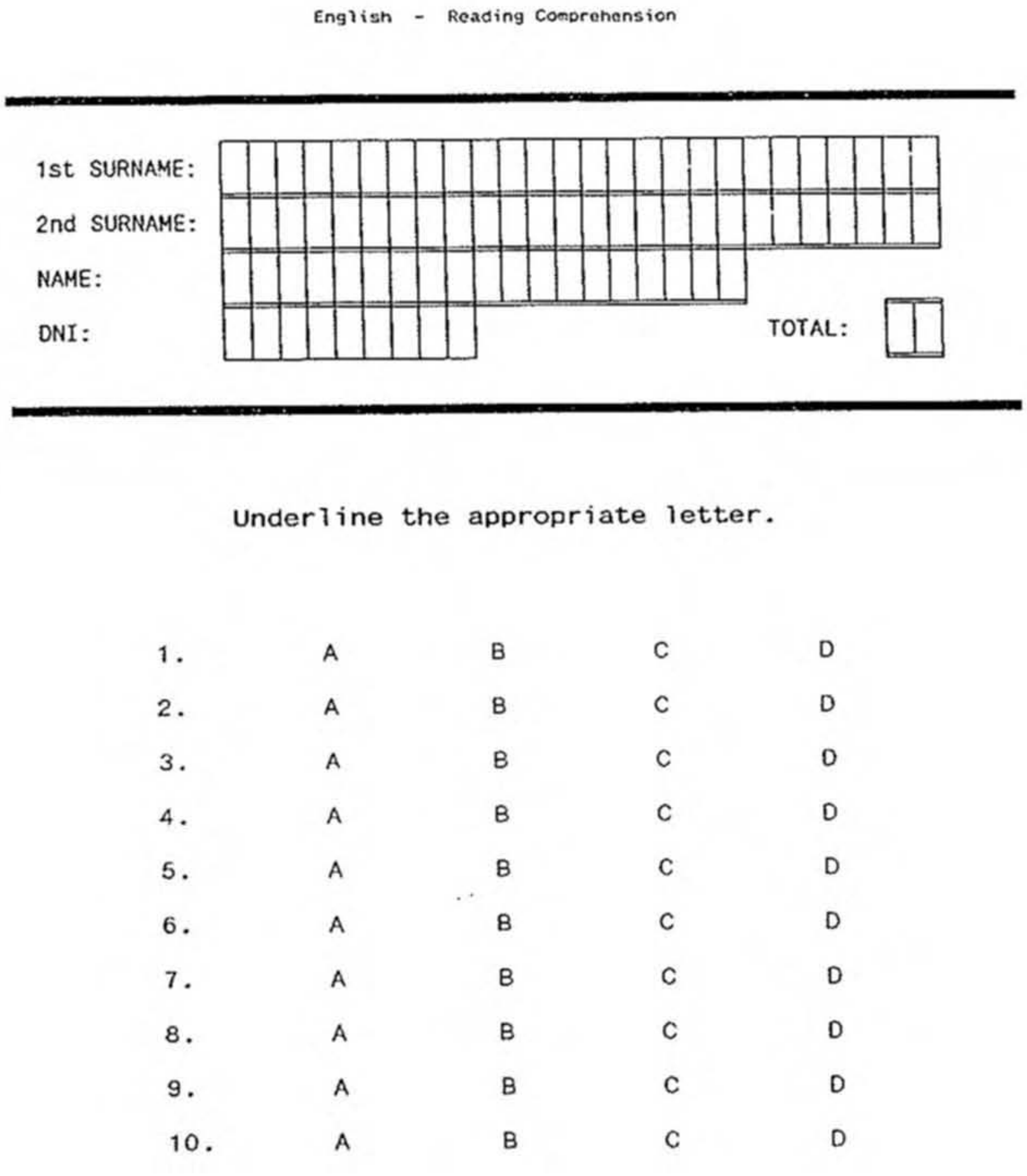
Nontheless there are signs that the Reykjavik intelligentsia is whipping up something of a backslap against the golden boy - some of its members have slated the book, calling it formulaic, catchpenny literature, and have complained that they're other icelandic writers more deserving of an international following.

As a novelist Olafsson has his floors. His stately style may not be to everyone's taste, and he lacks lyricism atmosphere is another week point - his anachronistic portrayal of pre-war Reykjavik cannot compare with the vivid way writers such like Einar Karason and Petur Gunnarsson have conjured up the post-war period.

He also resorts to obvious ploys to whet the readers appetite. But overall the book works extremely well both as a psycological thriller and as a character study.

Though *Absolution* is set for publication in Germany, France and Norway this autumn, there is little chance of Olafsson giving up his day job. He is saying he does not want to have to write to pay.

EXAMPLE: weak



English - Reading Comprehension

1st SURNAME:

2nd SURNAME:

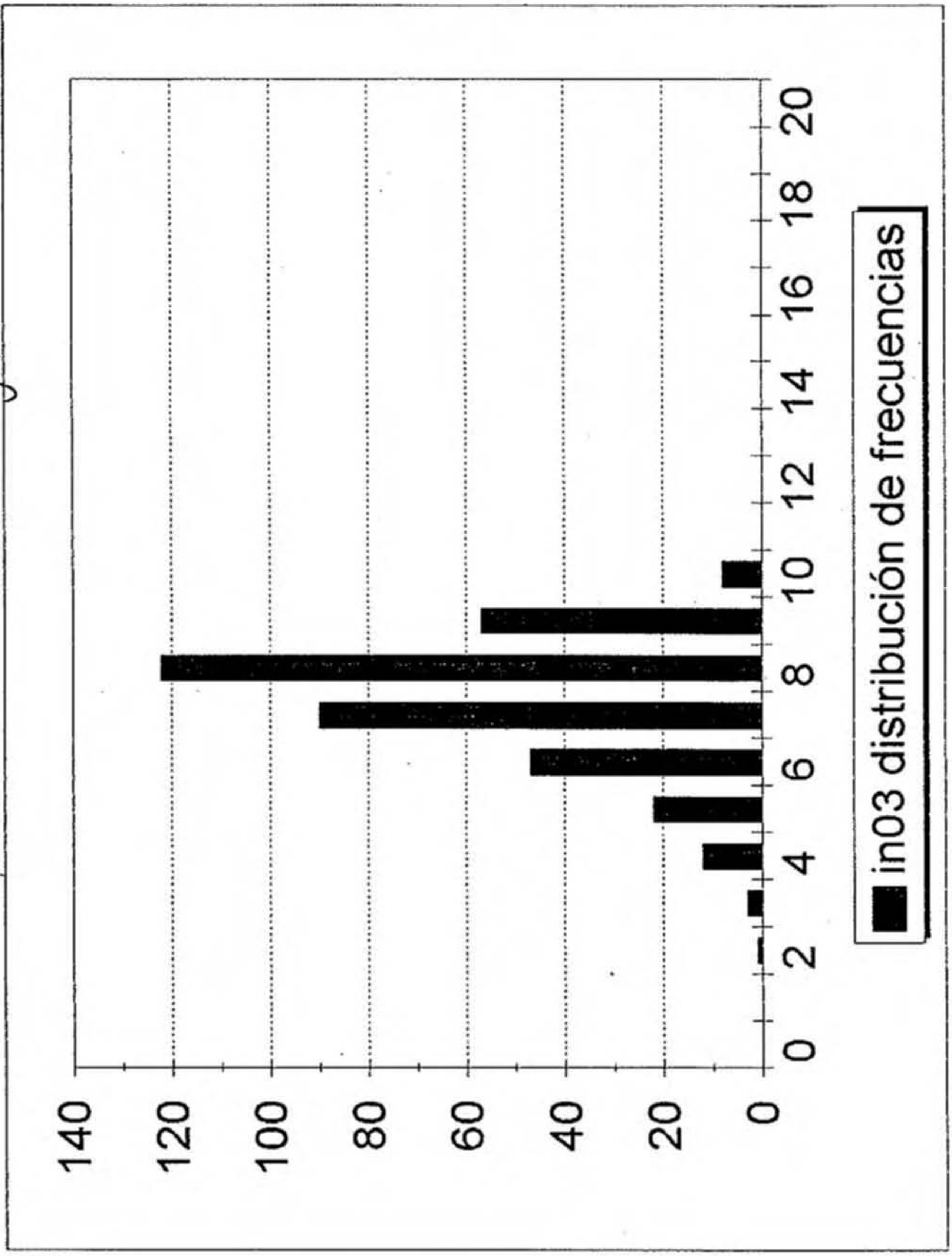
NAME:

DNI:

TOTAL:

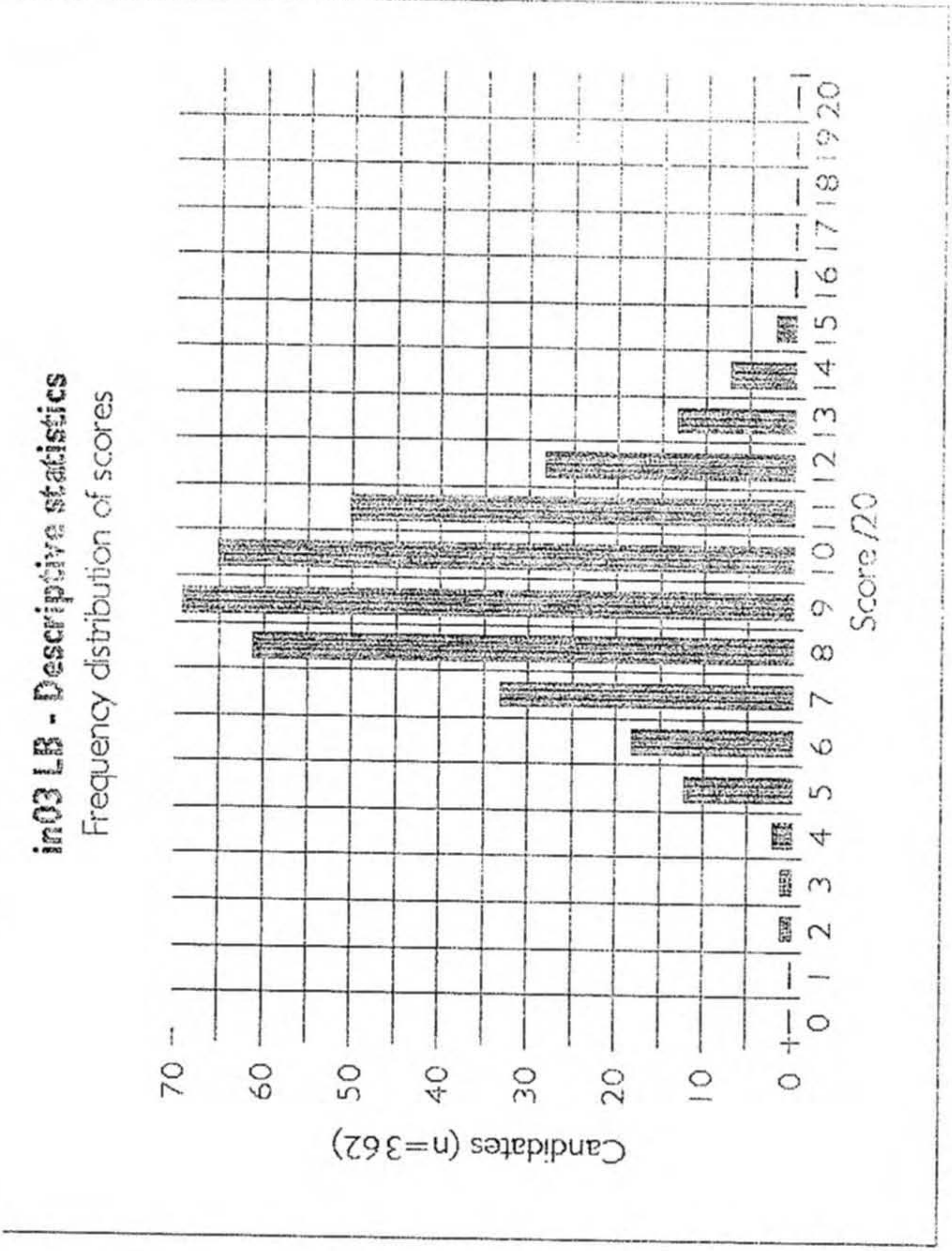
Underline the appropriate letter.

1.	A	B	C	D
2.	A	B	C	D
3.	A	B	C	D
4.	A	B	C	D
5.	A	B	C	D
6.	A	B	C	D
7.	A	B	C	D
8.	A	B	C	D
9.	A	B	C	D
10.	A	B	C	D



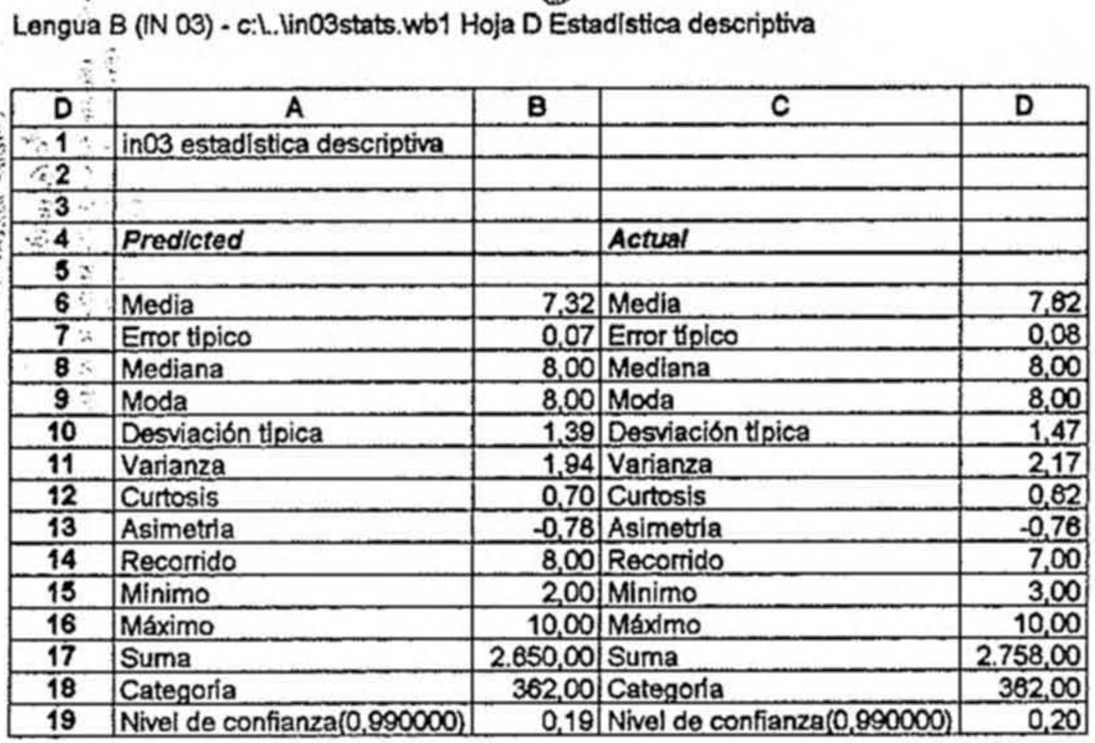
nc Ø mly

c:\document\thesis\results\stats\in03\in03stat.wb1



in03 LB - Descriptive statistics

Frequency distribution of scores



Lengua B (IN 03) - c:\..\in03stats.wb1 Hoja D Estadística descriptiva

D	A	B	C	D
1	in03 estadística descriptiva			
2				
3				
4	*Predicted*		*Actual*	
5				
6	Media	7,32	Media	7,82
7	Error típico	0,07	Error típico	0,08
8	Mediana	8,00	Mediana	8,00
9	Moda	8,00	Moda	8,00
10	Desviación típica	1,39	Desviación típica	1,47
11	Varianza	1,94	Varianza	2,17
12	Curtosis	0,70	Curtosis	0,82
13	Asimetría	-0,78	Asimetría	-0,76
14	Recorrido	8,00	Recorrido	7,00
15	Mínimo	2,00	Mínimo	3,00
16	Máximo	10,00	Máximo	10,00
17	Suma	2.650,00	Suma	2.758,00
18	Categoría	362,00	Categoría	362,00
19	Nivel de confianza(0,990000)	0,19	Nivel de confianza(0,990000)	0,20

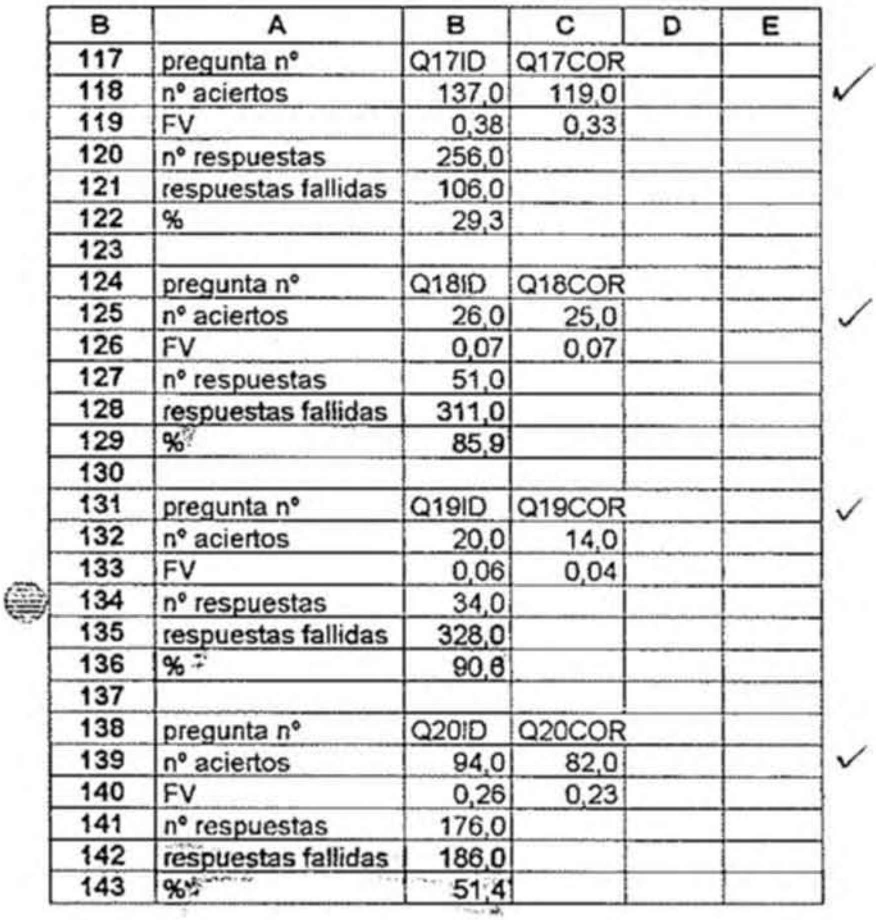
MCQ items

Prueba Julio '94 - Lengua B (IN 03) - c:\..\in03stats.wb1 Hoja B Valor fv

362

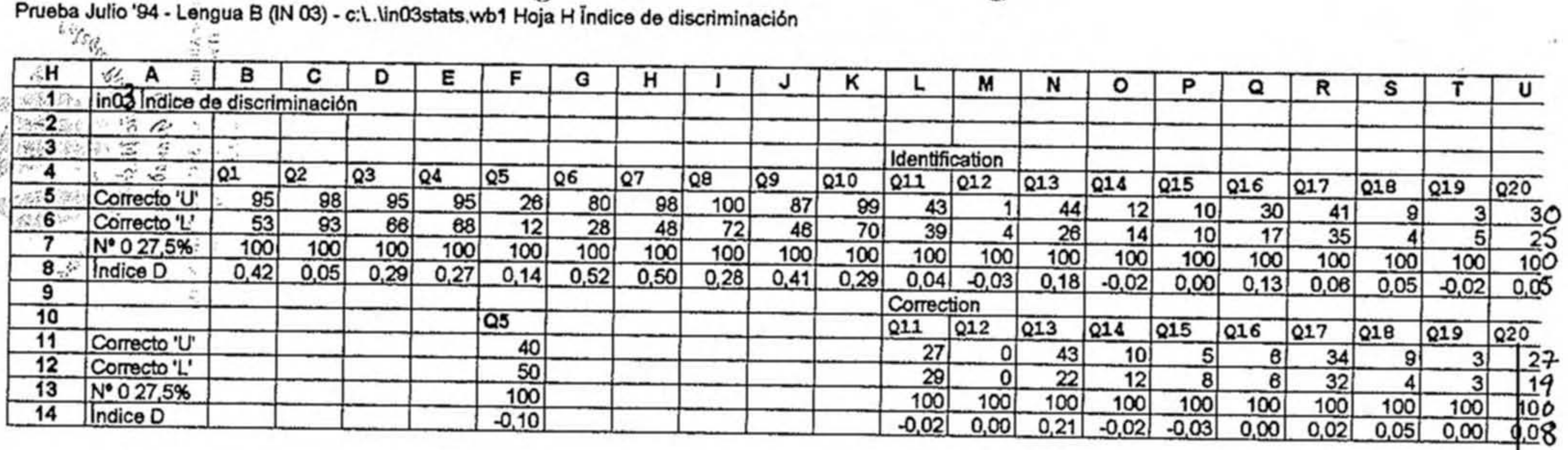
B	A	B	C	D	E	
1	in02fv					
2	Población =	362				
3						
4						
5	pregunta nº	Q1A	Q1B	Q1C	Q1D	
6	nº aciertos	265	8	62	24	✓
7	FV	0,73	0,02	0,17	0,07	
8	nº respuestas	359				
9	respuestas fallidas	3				
10	%	0,8				
11						
12	pregunta nº	Q2A	Q2B	Q2C	Q2D	
13	nº aciertos	6	350	3	4	✓
14	FV	0,02	0,97	0,01	0,01	
15	nº respuestas	363				
16	respuestas fallidas	-1				
17	%	-0,3				
18						
19	pregunta nº	Q3A	Q3B	Q3C	Q3D	
20	nº aciertos	4	299	14	43	✓
21	FV	0,01	0,83	0,04	0,12	
22	nº respuestas	360				
23	respuestas fallidas	2				
24	%	0,6				
25						
26	pregunta nº	Q4A	Q4B	Q4C	Q4D	
27	nº aciertos	14	1	309	39	✓
28	FV	0,04	0,00	0,85	0,11	
29	nº respuestas	363				
30	respuestas fallidas	-1				
31	%	-0,3				
32						
33	pregunta nº	Q5A	Q5B	Q5C	Q5D	
34	nº aciertos	56	60	164	67	✓
35	FV	0,15	0,17	0,45	0,19	
36	nº respuestas	347				
37	respuestas fallidas	15				
38	%	4,1				
39						
40	pregunta nº	Q6A	Q6B	Q6C	Q6D	
41	nº aciertos	10	124	33	179	✓
42	FV	0,03	0,34	0,09	0,49	
43	nº respuestas	346				
44	respuestas fallidas	16				
45	%	4,4				
46						
47	pregunta nº	Q7A	Q7B	Q7C	Q7D	
48	nº aciertos	52	283	5	21	
49	FV	0,14	0,78	0,01	0,06	
50	nº respuestas	361				
51	respuestas fallidas	1				
52	%	0,3				
53						
54	pregunta nº	Q8A	Q8B	Q8C	Q8D	
55	nº aciertos	4	329	4	25	
56	FV	0,01	0,91	0,01	0,07	
57	nº respuestas	[illegible]				
58	respuestas fallidas	0				
59	%	0,0				
60						

B	A	B	C	D	E	
59	%	0,0				
60						
61	pregunta nº	Q9A	Q9B	Q9C	Q9D	
62	nº aciertos	44	25	258	35	✓
63	FV	0,12	0,07	0,71	0,10	
64	nº respuestas	366				
65	respuestas fallidas	-4				
66	%	-1,1				
67						
68	pregunta nº	Q10A	Q10B	Q10C	Q10D	
69	nº aciertos	5	0	322	37	✓
70	FV	0,01	0,00	0,89	0,10	
71	nº respuestas	364				
72	respuestas fallidas	-2				
73	%	-0,6				
74						
75	pregunta nº	Q11ID	Q11COR			
76	nº aciertos	137,0	99,0			✓
77	FV	0,38	0,27			
78	nº respuestas	236,0				
79	respuestas fallidas	126,0				
80	%	34,8				
81						
82	pregunta nº	Q12ID	Q12COR			✓
83	nº aciertos	9,0	0,0			
84	FV	0,02	0,00			
85	nº respuestas	9,0				
86	respuestas fallidas	353,0				
87	%	97,5				
88						
89	pregunta nº	Q13ID	Q13COR			✓
90	nº aciertos	125,0	115,0			
91	FV	0,35	0,32			
92	nº respuestas	240,0				
93	respuestas fallidas	122,0				
94	%	33,7				
95						✓
96	pregunta nº	Q14ID	Q14COR			
97	nº aciertos	46,0	40,0			
98	FV	0,13	0,11			
99	nº respuestas	86,0				
100	respuestas fallidas	276,0				
101	%	76,2				
102						
103	pregunta nº	Q15ID	Q15COR			
104	nº aciertos	41,0	23,0			✓
105	FV	0,11	0,06			
106	nº respuestas	64,0				
107	respuestas fallidas	298,0				
108	%	82,3				
109						
110	pregunta nº	Q16ID	Q16COR			
111	nº aciertos	97,0	32,0			
112	FV	0,27	0,09			
113	nº respuestas	129,0				
114	respuestas fallidas	233,0				
115	%	64,4				
116						



Prueba Julio '94 - Lengua B (IN 03) - c:\..\in03stats.wb1 Hoja B Valor fv

B	A	B	C	D	E	
117	pregunta nº	Q17ID	Q17COR			
118	nº aciertos	137,0	119,0			✓
119	FV	0,38	0,33			
120	nº respuestas	256,0				
121	respuestas fallidas	106,0				
122	%	29,3				
123						
124	pregunta nº	Q18ID	Q18COR			
125	nº aciertos	26,0	25,0			✓
126	FV	0,07	0,07			
127	nº respuestas	51,0				
128	respuestas fallidas	311,0				
129	%	85,9				
130						
131	pregunta nº	Q19ID	Q19COR			✓
132	nº aciertos	20,0	14,0			
133	FV	0,06	0,04			
134	nº respuestas	34,0				
135	respuestas fallidas	328,0				
136	%	90,6				
137						
138	pregunta nº	Q20ID	Q20COR			
139	nº aciertos	94,0	82,0			✓
140	FV	0,26	0,23			
141	nº respuestas	176,0				
142	respuestas fallidas	186,0				
143	%	51,4				



Prueba Julio '94 - Lengua B (IN 03) - c:\..\in03stats.wb1 Hoja H Índice de discriminación

H	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	in03 Índice de discriminación																				
2																					
3												Identification									
4		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
5	Correcto 'U'	95	98	95	95	26	80	98	100	87	99	43	1	44	12	10	30	41	9	3	30
6	Correcto 'L'	53	93	66	68	12	28	48	72	46	70	39	4	26	14	10	17	35	4	5	25
7	Nº 0 27,5%	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
8	Índice D	0,42	0,05	0,29	0,27	0,14	0,52	0,50	0,28	0,41	0,29	0,04	-0,03	0,18	-0,02	0,00	0,13	0,06	0,05	-0,02	0,05
9												Correction									
10						Q5						Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
11	Correcto 'U'					40						27	0	43	10	5	6	34	9	3	27
12	Correcto 'L'					50						29	0	22	12	8	6	32	4	3	19
13	Nº 0 27,5%					100						100	100	100	100	100	100	100	100	100	100
14	Índice D					-0,10						-0,02	0,00	0,21	-0,02	-0,03	0,00	0,02	0,05	0,00	0,08

in03 Summary table: FV and D				
Item	FV	DI		
1	0.73	0.42	A	DATA:I5
2	0.97	0.05	A	DATA:O5
3	0.83	0.29	A	DATA:T5
6	0.49	0.52	A	DATA:AK5
10	0.89	0.29	A	Data:BD5
11	0.38	0.04	A	Data:BG8
12	0.02	-0.03	A	Data:BH8
13	0.35	0.18	A	Data:BI8
15	0.11	0.00	A	Data:BK8
16	0.27	0.13	A	Data:BL8
19	0.06	-0.02	A	Data:BO8
23	0.32	0.21	A	Data:BS7
26	0.09	0.00	A	DATA:BV
28	0.07	0.05	A	DATA:bX
30	0.23	0.08	A	DATA:BZ
4	0.85	0.27	B	data:z
5	0.45	0.14	B	data:ac
7	0.73	0.50	B	data:an
8	0.91	0.28	B	data:as
9	0.71	0.41	B	data:ay
14	0.13	-0.02	B	data:bj
17	0.38	0.06	B	data:bm
18	0.07	0.05	B	data:bn
20	0.26	0.05	B	data:bp
21	0.27	-0.02	B	data:bq
22	0.00	0.00	B	data:br
24	0.11	-0.02	B	data:bt
25	0.06	-0.03	B	data:bu
27	0.33	0.02	B	data:bw
29	0.04	0.00	B	data:by

d.