

# Spanish



## INDEX

Introduction to DIALANG	2
Information about the Dialang Items	6
Dialang Spanish Reading Comprehension Items	8
Dialang Spanish Listening Comprehension Items	13



## Introduction to DIALANG

### Context

DIALANG is a computer-based and computer-scored suite of diagnostic tests of foreign language abilities in 14 European languages. There are tests of Reading, Writing, Listening, Structures and Vocabulary, in Danish, Dutch, English, Finnish, French, German, Greek, Icelandic, Irish, Italian, Norwegian, Portuguese, Spanish and Swedish. In addition, there is a Vocabulary Size Placement Test for each language, and, also for each language, there are self-assessment 'I Can' statements for Reading, Writing and Listening. DIALANG provides extensive feedback to users, and advice on how they might improve their abilities.

The DIALANG Project was funded by the European Commission under the Socrates LINGUA programme, ACTION D. The initial development of DIALANG, beginning at the end of 1996, was the result of collaboration among over 20 partner institutions, most of them universities, throughout Europe, coordinated by the University of Jyväskylä. For full details, visit the DIALANG website at [www.dialang.org](http://www.dialang.org). The piloting, calibration and further development of the system was coordinated by the Free University of Berlin. All tests can be downloaded from the DIALANG website, free of charge, and can be taken over the Internet, again free of charge, from any computer on which the program is installed.

DIALANG is intended to help learners diagnose their strengths and weaknesses in any or all of the five abilities tested, to encourage learners to self-assess their abilities, and to compare their self-assessment with their test results. By so doing it is intended to enhance learner autonomy by encouraging learner reflection on what they know and can do in their target foreign language, and thereby to increase their awareness of what is involved in learning or improving in a foreign language.

The DIALANG Project began in 1996, when the first draft of the Council of Europe's Common European Framework of Reference (CEFR) was already available, and so it was possible to base the DIALANG Assessment Framework and Specifications on the CEFR. DIALANG also used the 'Can Do' statements contained in the CEFR, but adjusted their wording to suit its self-assessment purposes, essentially by changing 'Can-Do' into 'I Can' and at times simplifying the language to make it more accessible to learners. In addition, extensive explanations of the feedback given by the DIALANG system were based on the CEFR scales, and the advice provided to learners on how to improve their ability to use the language used the information contained in the CEFR's Descriptive Scheme (which became Chapters 4, 5, 6, and 7 in the 2001 publication - Council of Europe, 2001).

DIALANG not only contains tests in its 14 languages, but it also enables users to read test instructions and rubrics, help facilities, introductions and, most importantly, all self-assessment statements, feedback and advice, in any of these 14 languages. Once agreement had been reached on the appropriate wording, particularly of the 'I Can' statements, and the Explanatory and Advisory Feedback, these were then translated from the original English into the 13 languages of DIALANG, and checked for the quality of the translations. Thus DIALANG is multilingual in both the tests available, and in the interface and administration of the system.

Although intended to offer detailed diagnosis of language skills and subskills (for full details of these, see Alderson, 2005), DIALANG also reports the learner's ability in the skill or aspect of language being tested, as a level on the CEFR (A1 to C2), rather than as a numerical score.

DIALANG is intended to contribute to life-long learning and learner autonomy, and thus is "aimed" at any adult learner of foreign languages, whether or not that learner is engaged in institution-based language learning. However, many institutions consider the DIALANG system to be of value in placing students into language classes which are based on the CEFR, and thus, although intended to be diagnostic, the tests are increasingly being used for placement purposes. At the time of writing, roughly 1,500 users take a DIALANG test every day.

## Test development process

### a) Content quality

The DIALANG tests were developed by Assessment Development Teams (ADTs) for each of the 14 languages. Typically a team was composed of language educationalists, usually at institutions of adult or higher education. Members were experienced teachers, researchers and assessment experts, where these existed for the language in question. In several cases, where the traditions of assessment were different from those espoused in DIALANG and the CEFR, familiarisation with the CEFR and with appropriate methods of test and item writing was made available and a number of training sessions were held. The tests themselves, as already noted, were based on the DIALANG Assessment Framework (DAF) and DIALANG Assessment Specifications (DAS), themselves based on or derived from the CEFR and related User Guides. Detailed Guidelines for Item Writers were also developed and made available as the formats (test methods) of the tests and an associated computer-based authoring system became available.

Test development was coordinated by two Test Development Coordinators, who were responsible for the issue and updating of all documentation and guidance, and who communicated with the ADTs through the ADT Team Leader. As items were developed, first on paper and then over the Internet using the Authoring software developed by the Project, they were subject to an extensive process of quality review, first by peers within the ADT, then by the Team Leader, and then by external reviewers commissioned by the Test Development Coordinators, using the Internet-based Item Review software. Once revisions had been made and reviewed by the ADT, the Test Development Coordinators made a second review to ensure that all items conformed to the DAF and DAS, and to the requirements of the computer programs. In this way, large pools of items were developed for each skill for each language. Finally, in preparation for empirical piloting, ADT Team leaders were requested to select the best 300 items from their own item pool for the first round of piloting.

### b) Piloting and calibration

In the second phase of the Project, arrangements were made for the computer-based piloting of items using the specially developed Pilot Tool software also developed by the Project (in the first phase of the Project, Finnish had been piloted in paper and pencil form, in order to try out both the overall pilot design and the calibration analyses). Initially a piloting design was implemented where 12 booklets would be piloted, containing 30 items from a skill (Reading, Writing, Listening) plus 20 items from a language aspect (Structures, Vocabulary), with each item appearing in two separate booklets, as well as 150 items from the Vocabulary Size Placement Test (VSPT) and between 30 and 40 self-assessment (SA) statements. Booklets were assigned randomly to pilot test-takers. This design would enable IRT-based item calibrations

to take place, using the computer program OPLM, once 600 learners had taken a pilot test (implying 100 responses per item) and a second calibration was to take place once 1200 users had responded (200 responses per item).

However, once sufficient numbers of responses to the VSPT and SAs had been gathered to enable calibrations, the pilot booklet design was modified to reduce the number of VSPT items being tested to 99 and the SA statements to 18 per skill. This allowed more test items to be included in each booklet, and only 450 test-takers would be required before an initial item calibration was possible. To date, initial calibrations have been possible for Finnish, English, French, German and Spanish, and Italian, Swedish and Dutch are getting close to the target number. English has undergone a second calibration, whose results were remarkably similar to the first round, suggesting a second calibration may not be necessary for other languages.

## Standard Setting

Two different methods of standard setting were developed in DIALANG, and the Project developed guidelines which were subsequently published in the Manual and the Supplement. Details are available in Alderson (2005). In essence, the first method was a modified Angoff procedure, where probability judgements by up to 12 experts were replaced by Yes-No decisions (numbers of judges varied by language and skill).

“The task for the judge is to state for each item in the test

*At what CEF level can a test taker already answer the following item correctly?”*

...

In practice the procedure follows a pattern also presented in the Manual:

- Familiarisation with the CEF
- Training (with feedback) on how to judge ... new items ... in terms of CEF levels;
- Judgement of items to be related to the CEF

(Manual, page 91)

This method was applied to Finnish, English and Spanish, and detailed results are reported in Alderson (2005).

Partly because of the lack of an adequate method of judging the reliability and consistency of judges, a second process was developed which involved experts in making two sets of judgements. The first round required them to place items into piles, representing the different levels of the CEFR, starting with A1, in answer to the question:

*“Is it reasonable to require that a learner at level A1 gets this item right?”*

If yes, then the item belongs to A1. If not, then it is a candidate for the next level up. Once piles have been created, they are checked for internal consistency and items may be moved if necessary.

After a period of time, minimally two to three hours, the process is repeated, thereby enabling intra-rater consistency to be calculated alongside the inter-rater agreement. Detailed results are available in Alderson (2005). Only reliable judges are retained in the subsequent analysis.

Finally, once item calibration details are available, the empirical results are combined with the judged CEFR level for each item, in a computer program specially developed by the Project.

## References

**Alderson, J. C.** (2005) *Diagnosing foreign language proficiency: the interface between learning and assessment*. London: Continuum Books

**Council of Europe** (2003) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF). Manual: Preliminary Pilot Version*. Strasbourg: Language Policy Division, Council of Europe.

## Information about the Dialang Items

The items presented in this section are taken from Version 1 of the freely available English test and represent a selection across the different CEFR levels. Each item is accompanied by a summary of its content taken from the Dutch CEF Grid ([www.ling.lancs.ac.uk/cefgrid](http://www.ling.lancs.ac.uk/cefgrid)), itself closely based on the CEFR. In addition each item is accompanied by information on its statistical properties.

### Statistical information

The calibration and standard setting processes described in the introduction to Dialang resulted in information about the properties of each item in terms of classical test theory - in particular, item facilities and discriminations. In addition, IRT analyses resulted in estimations of the ability of learners getting a given score, in terms of theta estimates of ability. Cut-offs on this theta scale are based on the standard-setting judgements, and an item's logit value can be related to the logit values for ability levels.

### Guideline To Information On DIALANG items

Below follows an explanation of the statistical and other information accompanying the selected DIALANG items:

- the Framework level of the item is indicated immediately before the screenshot of the item
- the subskill of listening that the item was designed to measure is indicated immediately after the screenshot; DIALANG listening items aim at tapping one of these three skills:
  - o Identifying the main idea(s) or information in the spoken discourse.
  - o Listening intensively for specific detail.
  - o Inferencing; i.e., the ability to infer meaning on the basis of spoken discourse.

The table under 'Information from piloting' describes the results for the booklets where the item was piloted.

- booklet number = the number of the pilot booklet in which the item was piloted (each DIALANG item appeared in two different booklets in the pilot design)
- n size = the number of pilot test takers for the booklet in question
- Facility = percent of pilot test takers who replied the item correctly
- Rpbis (unweighted) = point-biserial correlation between the item and the other listening items in the pilot booklet when all items have equal weight
- Rpbis (weighted) = point-biserial correlation between the item and the other listening items in the pilot booklet when the items are weighted by their discrimination index ('A' in the table under 'Results of calibration')

The table 'Results of calibration' display the key statistical information for the item from the OPLM analysis.

nr = number of the item in the OPLM analysis

label = label of the item (which in fact is also the item ID in the DIALANG system)

A = discrimination index

B = theta value of the item (which indicates its difficulty / position on the scale after the items have been scaled; the cut-offs for the Framework levels are expressed on this same scale)

SE(B) = standard error of the theta value

S = S test; in this the respondents are grouped into two or more (max 8) groups of equal size and the ability of the item to discriminate the groups is analysed

DF = degrees of freedom in the S analysis

P = probability of the S test (this in fact indicates whether the item studied fits the overall model)

M, M2, M3 = other indicators of the model fit for the item

— M = in this analysis, two groups are formed of the test takers and it is analysed whether the item can discriminate the two groups (the low group are those whose probability to get the item right is at most .4, whereas in the high group the probability is .6 or higher)

— M2 = in this analysis, respondents are split into two halves according to their overall score in the test / booklet

— M3 = in this analysis, the respondents are grouped into three groups of equal size, according to their test score

P or M-value of 99.999 indicates that a proper value cannot be computed because the respondents cannot be divided into equal groups; this occurs with very easy or very difficult items. For the same reason, the S test value can sometimes be .000.

An M value of the size of + or - 2 or above indicates that there may be a problem in the item (it cannot properly discriminate test takers at different ability levels)

## Analysis of item content

The items on this CD have been analysed using the Dutch CEF Grid produced by the Dutch CEF Construct Project. This Project aimed to help test developers and other language educationists construct or relate test items to the CEFR. (The Final Report of the Project has been included on this CD for ease of reference.) One major outcome of the Project was an Internet-based Grid which can be used to help characterise reading and listening texts, items and tasks. Application of the Grid results in summary tables and the Grid has been used with the DIALANG items to produce the summary tables below. (Readers of this Users' Guide can access the Grid at:

[www.ling.lancs.ac.uk/cefgrid](http://www.ling.lancs.ac.uk/cefgrid)

and can use it in the content analysis of their own texts, items and tasks.)

## Dialang Spanish Reading Comprehension Items

### General information

There were 87 – 92 pilot test takers per each reading item in the OPLM analyses. Overall data – model fit was satisfactory ( $p = .25$ ).

The cut-offs on the theta scale for the CoE levels are based on standard setting procedures where 12 experts judged the items; the standard setting data were analysed with a special programme designed by Norman Verhelst at CITO.

### Cut-off points for the Framework levels for the DIALANG German reading items:

under A1: theta under $-.63$
A1: theta between $-.63$ and $-.32$
A2: theta between $-.32$ and $+.23$
B1: theta between $+.23$ and $+.60$
B2: theta between $+.60$ and $+.77$
C1: theta between $+.77$ and $+.91$
C2: theta above $+.91$



## A2

Revisión de ejercicios v1.14 [ONLINE] Q 002608

Á á É é Í í Ó ó Ú ú Ü ü Ñ ñ ¿ i

Lea el texto y complete la tarea llenando la(s) casilla(s) vacía(s). Haga clic en la casilla para que aparezca una lista de opciones. Escoja su respuesta haciendo clic en ella.

Para conseguir los benéficos efectos de la meditación hay que dedicarle al menos 20 minutos diarios, antes o después de la jornada laboral.

- Medita en un lugar silencioso y con una temperatura templada.
- Ponte una ropa cómoda y holgada.
- No practiques nunca con el estómago lleno.
- Elige una postura cómoda: sentado en el suelo con la espalda recta y relajada, o en una silla.
- Cierra los ojos o concéntrate en un punto fijo del suelo.
- Concéntrate en tu propia respiración sin preocuparte si la realizas bien o mal, sólo intenta seguir su ritmo mecánicamente.
- No opongas resistencia a los pensamientos que lleguen. Poco a poco verás como la mente aprenderá a quedarse en blanco.

**Lea el artículo y marque la respuesta en la caja**

Hay que practicar la meditación

- después de comer
- en el lugar más cálido de la casa
- vestido cómodamente
- durante la jornada laboral

Ayuda    Siguiente    Skip

## Content analysis

### Text Characteristics:

Task:	2608
1. Text Source:	Instructional material
2. Authenticity:	Authentic
3. Discourse type:	Mainly Instructive
4. Domain:	Personal
5. Topic:	7. Health and bodycare
6. Nature of Content:	mostly concrete content
7. Text Length:	114
8. Vocabulary:	rather extended
9. Grammar:	mainly simple structures
Comprehensible by learner at CEF level:	A2/B1

### Item Characteristics:

Item:	1
14. Item Type:	1. Multiple choice
15. Operations:	Recognise and Retrieve
15. Operations:	Explicit
15. Operations:	Detail
Item Level Estimated:	A2/B1

## Statistical analysis

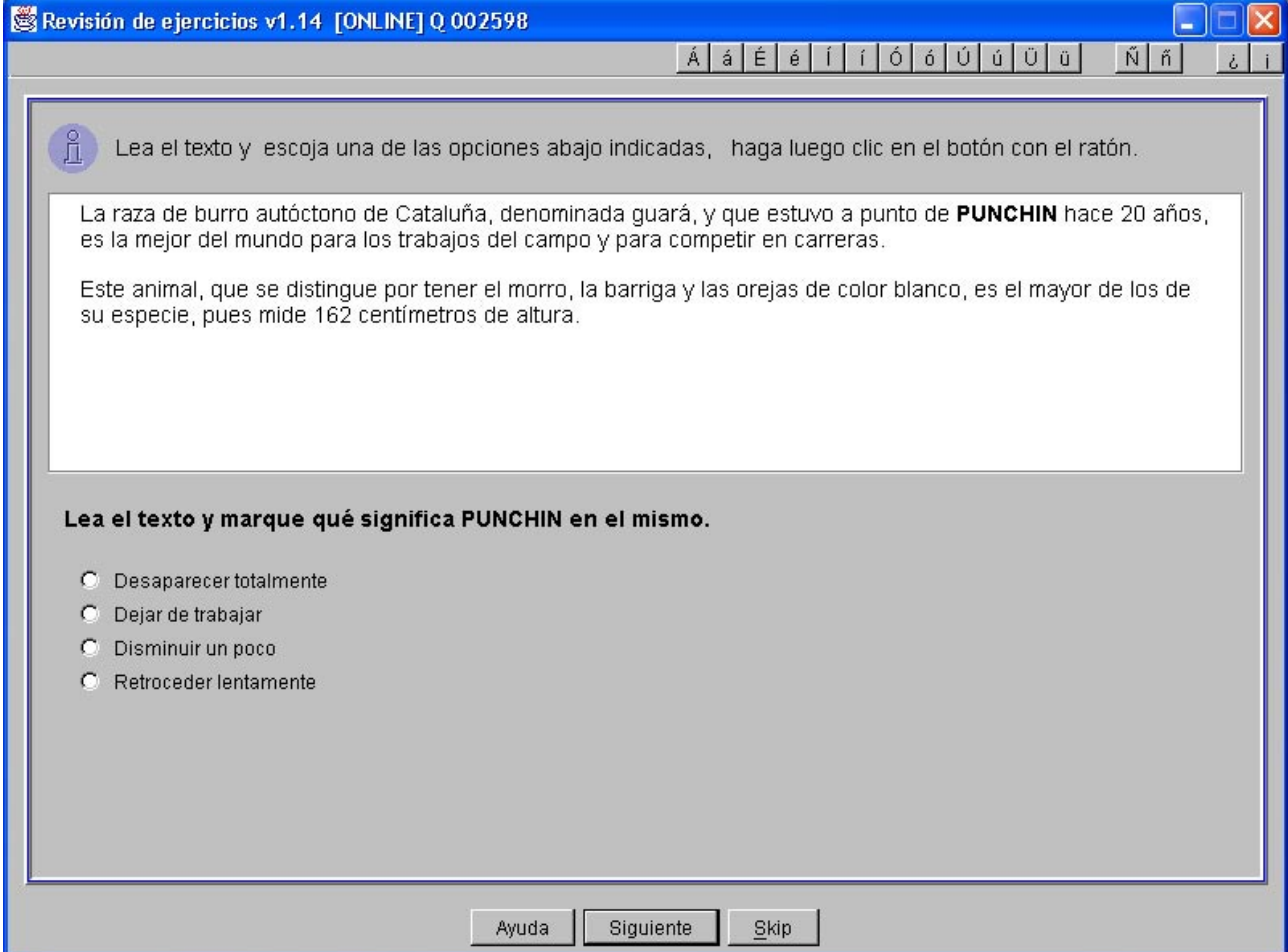
### Information from piloting:

Booklet number	n size	Facility	Rpbis (unweighted)	Rpbis (weighted)
3	43	.744	.719	.747
4	44	.705	.598	.600

### Results of calibration:

nr label	A	B	SE(B)	S	DF	P	M	M2	M3
109 R002608	4	-.001	.079	.580	1	.446	-1.043	-.426	-.130

## A2



Revisión de ejercicios v1.14 [ONLINE] Q 002598

Á á É é Í í Ó ó Ú ú Ü ü Ñ ñ ¿ ¡

Lea el texto y escoja una de las opciones abajo indicadas, haga luego clic en el botón con el ratón.

La raza de burro autóctono de Cataluña, denominada guará, y que estuvo a punto de **PUNCHIN** hace 20 años, es la mejor del mundo para los trabajos del campo y para competir en carreras.

Este animal, que se distingue por tener el morro, la barriga y las orejas de color blanco, es el mayor de los de su especie, pues mide 162 centímetros de altura.

Lea el texto y marque qué significa **PUNCHIN** en el mismo.

- Desaparecer totalmente
- Dejar de trabajar
- Disminuir un poco
- Retroceder lentamente

Ayuda Siguiente Skip

## Content analysis

### Text Characteristics:

Task:	2598
1. Text Source:	Reference books
2. Authenticity:	Authentic
3. Discourse type:	Mainly Descriptive
4. Domain:	Public
5. Topic:	15. Other
Specify other topic:	animals
6. Nature of Content:	mostly concrete content
7. Text Length:	65
8. Vocabulary:	mostly frequent vocabulary
9. Grammar:	mainly simple structures
Comprehensible by learner at CEF level:	B1

### Item Characteristics:

Item:	1
14. Item Type:	1. Multiple choice
15. Operations:	Make inferences
15. Operations:	Explicit
15. Operations:	Detail
Item Level Estimated:	B1

*Comment: The expert judges doing the content analysis overestimated the difficulty level of the item.*

## Statistical analysis

### Information from piloting:

Booklet number	n size	Facility	Rpbis (unweighted)	Rpbis (weighted)
2	46	.652	.650	.647
3	43	.535	.673	.679

### Results of calibration:

nr label	A	B	SE(B)	S	DF	P	M	M2	M3
92 R002598	4	.218	.071	.003	1	.957	.230	-.663	-1.091

## Dialang Spanish Listening Comprehension Items

### General information

There were 75 – 98 pilot test takers per each reading item in the OPLM analyses. Overall data – model fit was satisfactory ( $p = .39$ ).

The cut-offs on the theta scale for the CoE levels are based on standard setting procedures where 12 experts judged the items; the standard setting data were analysed with a special programme designed by Norman Verhelst at CITO.

### Cut-off points for the Framework levels for the DIALANG German LISTENING items:

under A1: theta under -.65
A1: theta between -.65 and -.11
A2: theta between -.1 and +.43
B1: theta between +.43 and +.85
B2: theta between +.85 and +1.66
C1: theta between +1.66 and +2.83
C2: theta above +2.83



**A1**

## Content analysis

### Text Characteristics:

Task:	3629
1. Text Source:	Telephone conversations
2. Authenticity:	Authentic
3. Discourse type:	Mainly Descriptive
4. Domain:	Personal
5. Topic:	3. Daily life
6. Nature of Content:	only concrete content
7. Text Length:	13"
8. Vocabulary:	only frequent vocabulary
9. Grammar:	only simple structures
10. Text Speed:	normal
11. No of participants:	two
12. Accent Standard:	Standard pronunciation
13. Clarity of Articulation:	normally articulated
14. How often played:	played once
Comprehensible by learner at CEF level:	A1/A2

### Item Characteristics:

Item:	1
14. Item Type:	1. Multiple choice
15. Operations:	Recognise and Retrieve
15. Operations:	Explicit
15. Operations:	Main idea/gist
Item Level Estimated:	A1/A2

## Statistical analysis

### Information from piloting:

Booklet number	n size	Facility	Rpbis (unweighted)	Rpbis (weighted)
11	52	.923	.616	.629
12	46	.957	.423	.446

### Results of calibration:

nr label	A	B	SE(B)	S	DF	P	M	M2	M3
53 L003629	5	-.549	.110	.000	0	99.999	-.169	-.349	-.573



**A1**



## Content analysis

### Text Characteristics:

Task:	2563
1. Text Source:	Interpersonal dialogues and conversations
2. Authenticity:	Authentic
3. Discourse type:	Mainly Narrative
4. Domain:	Personal
5. Topic:	3. Daily life
6. Nature of Content:	only concrete content
7. Text Length:	17"
8. Vocabulary:	only frequent vocabulary
9. Grammar:	only simple structures
10. Text Speed:	normal
11. No of participants:	one
12. Accent Standard:	Standard pronunciation
13. Clarity of Articulation:	normally articulated
14. How often played:	played once
Comprehensible by learner at CEF level:	A1/A2

### Item Characteristics:

Item:	1
14. Item Type:	1. Multiple choice
15. Operations:	Recognise and Retrieve
15. Operations:	Explicit
15. Operations:	Detail
Item Level Estimated:	A1/A2

## Statistical analysis

### Information from piloting:

Booklet number	n size	Facility	Rpbis (unweighted)	Rpbis (weighted)
9	37	.892	.476	.486
12	46	.891	.512	.507

### Results of calibration:

nr label	A	B	SE(B)	S	DF	P	M	M2	M3
15 L002563	4	-.485	.104	.000	0	99.999	.303	-.652	-.191

## Content analysis



**A1**

**Text Characteristics:**

Task:	2548
1. Text Source:	Telephone conversations
2. Authenticity:	Authentic
3. Discourse type:	Mainly Instructive
4. Domain:	Personal
5. Topic:	3. Daily life
6. Nature of Content:	only concrete content
7. Text Length:	8"
8. Vocabulary:	only frequent vocabulary
9. Grammar:	only simple structures
10. Text Speed:	normal
11. No of participants:	two
12. Accent Standard:	Standard pronunciation
13. Clarity of Articulation:	normally articulated
14. How often played:	played once
Comprehensible by learner at CEF level:	A1/A2

**Item Characteristics:**

Item:	1
14. Item Type:	1. Multiple choice
15. Operations:	Recognise and Retrieve
15. Operations:	Explicit
15. Operations:	Detail
Item Level Estimated:	A1/A2

**Statistical analysis****Information from piloting:**

Booklet number	n size	Facility	Rpbis (unweighted)	Rpbis (weighted)
9	37	.811	.513	.549
10	38	.763	.372	.348

**Results of calibration:**

nr label	A	B	SE(B)	S	DF	P	M	M2	M3
26 L002548	3	-.348	.108	.000	0	99.999	1.022	-.173	.356

**Content analysis**



**A1**

**Text Characteristics:**

Task:	3771
1. Text Source:	Interpersonal dialogues and conversations
2. Authenticity:	Pedagogic
3. Discourse type:	Mainly Descriptive
4. Domain:	Personal
5. Topic:	1. Personal identification
6. Nature of Content:	only concrete content
7. Text Length:	27"
8. Vocabulary:	only frequent vocabulary
9. Grammar:	only simple structures
10. Text Speed:	normal
11. No of participants:	two
12. Accent Standard:	Standard pronunciation
13. Clarity of Articulation:	normally articulated
14. How often played:	played once
Comprehensible by learner at CEF level:	A1/A2

**Item Characteristics:**

Item:	1
14. Item Type:	7. Short answer question
15. Operations:	Recognise and Retrieve
15. Operations:	Explicit
15. Operations:	Detail
Item Level Estimated:	A1

**Statistical analysis****Information from piloting:**

Booklet number	n size	Facility	Rpbis (unweighted)	Rpbis (weighted)
11	52	.865	.468	.492
12	46	.870	.375	.432

**Results of calibration:**

nr label	A	B	SE(B)	S	DF	P	M	M2	M3
57 L003771	3	-.439	.113	.000	0	99.999	.245	.256	-.007

**Content analysis**



**B1**

**Text Characteristics:**

Task:	3642
1. Text Source:	Public speeches, lectures, presentations, sermons
2. Authenticity:	Authentic
3. Discourse type:	Mainly Narrative
4. Domain:	Public
5. Topic:	15. Other
Specify other topic:	Arts
6. Nature of Content:	mostly concrete content
7. Text Length:	31"
8. Vocabulary:	mostly frequent vocabulary
9. Grammar:	mainly simple structures
10. Text Speed:	normal
11. No of participants:	two
12. Accent Standard:	Standard pronunciation
13. Clarity of Articulation:	clearly articulated
14. How often played:	played once
Comprehensible by learner at CEF level:	B1

**Item Characteristics:**

Item:	1
14. Item Type:	7. Short answer question
15. Operations:	Recognise and Retrieve
15. Operations:	Explicit
15. Operations:	Detail
Item Level Estimated:	B1

**Statistical analysis****Information from piloting:**

Booklet number	n size	Facility	Rpbis (unweighted)	Rpbis (weighted)
11	52	.481	.487	.530
12	46	.348	.634	.641

**Results of calibration:**

nr label	A	B	SE(B)	S	DF	P	M	M2	M3
54 L003642	4	.508	.067	.022	1	.881	-.243	-.844	.061