Test Production and Test Delivery Standards and Procedures

Contents

1. Introduction

This document provides an overview of the test production process. It describes the steps of the process, the responsible parties and refers to the documentation behind the process. It goes through the following stages of the process of item and task preparation:

1. The management of the process
2. Writing test materials
3. Quality and version control
4. Test delivery
5. Marking
6. Statistical analysis
7. Item banking

Test materials are written according to the detailed test specifications by a team of trained item writers. Material writing is commissioned on the basis of a running exam calendar, always ensuring that an ample period is allotted to quality control before administration.

The process of item writing is built up of numerous steps and stages and relies on internal staff as well as a number of experts from different fields (See Appendix 1: The steps of the test production and delivery process). The documentation and steps of the procedures used to commission, write, check, and pre-test items, and the statistics used for result determination and item banking are described in this document.

2. Item and task preparation

2.1 The management of the process

The Internal Assessement Team coordinates the writing process. The logistics of the process is coordinated by the Test Production Coordinator, and professionally supervised and managed by the Test Production Manager with the contribution of the members of the Assessment Team.

Based on the *Item Writing Schedule*, dates and responsible parties are allocated to each step of the paper production. The schedule is managed by the Test Production Coordinator and updated after each step.

2.2 Writing test materials

2.2.1 Item writers

Test items and tasks are produced by trained, professional item writers. Item writers are recruited based on their qualifications and experience relevant to the subject.

Item writers are commissioned individual tasks or complete sections of a paper based on the annual plan as it appears in the *Item Writing Schedule*. Writers are required to produce their items and tasks according to the detailed instructions in the *Item Writer's Guide* appropriate to the level of the target examination.

2.2.2 Test materials

The *Item Writer's Guide* is the scheme all writers have to refer to. Among other information, the IWG gives details about the following:

1. test specification
2. task types
3. text types
4. topics
5. the language focus for each level
6. sample tasks

The specifications ensure that the items or tasks written by different writers are consistent regarding difficulty, reliability and validity. The item writers must always use the level specific

templates and checklists for preparing tasks.

The preliminary step of producing materials is commissioning. Materials are commissioned from either internal or external item writers, who should observe the following requirements:

1. use the templates and checklists provided
2. follow copyright regulations for the use of pictures and illustrations
3. provide detailed descriptions for illustrators in connection with visual prompts
4. follow the file naming convention of the items

The flowchart are provided as an appendix to this document (Appendix 2: Flowchart).

2.3 Internal materials review and editing

2.3.1 Technical and content check

The first step of the process is the submission of draft items (**Step 1**). The submitted test materials must be checked for quality and consistency. After the submission of the preliminary versions of test items or tasks, a member of the Assessment Team checks the items against the IWG (**Step 2**). This is a technical check (concerning wordcount, format and accuracy) and a content check (concerning level, relevance and appropriacy) at the same time. The item writers are provided feedback on the basis of which they can make the necessary amendments (**Step 3**).

2.3.2 First Moderation

As soon as the assessment team receives the amended items, a second technical and content check is performed (**Step 4**), which is followed by an expert panel review (moderation) (**Step 5**). The members of the panel (trained EFL experts, markers, teachers, linguists and item writers) moderate (edit) the tasks and answer keys. The panel consists of a minimum of four members including a co-ordinator, who schedules and documents the process.

The panel – using expert judgement – considers the following criteria:

1. relevance of topic
2. content accuracy
3. format
4. clarity and wording

5. answer keys
6. instructions
7. coverage of different topic areas
8. CEFR level

The moderators also decide on the graphics and instructions to illustrators, finalise the listening scripts and give instructions to voice actors. Items are usually amended by the moderators following the panel review session; however, in case of quality issues, the tasks may be returned to the writer for rewriting according to feedback given, or rejected, with a reason.

After moderation, the Test Production Coordinator commissions the graphics and the recordings (**Step 6-7**), and makes sure the material is ready for the deadline. Once the illustrator has provided the graphics, the Test Production Coordinator adds the graphics, checks the quality, content, etc. and prepares the paper for pre-testing (**Step 8-9**).

2.3.3 Materials try-out

Before test materials are incorporated into live exams, some form of try-out is needed so that test taker responses can be observed and analysed. The aim is to gain data from groups of people who are potential test takers. Consequently, it is ensured that all regions with a significant test taking population are represented at this stage of the process. The forms of try-out used in the process are piloting and pre-testing (**Step 10**).

2.3.3.1 Piloting

Piloting involves a small number of participants who complete the test, respond to questions and comment on the test material. The tasks and participants are chosen by the Research and Development Team, who also analyse the responses and make recommendations to the Assessment Team in connection with the test items. The responses of the participants are analysed with qualitative methods.

2.3.3.2 Pre-testing

Pre-testing is conducted to see how objectively marked items work with test takers who are similar to the expected test taker population. Pre-testing is performed under live testing conditions for which an appropriate sample size of 50-100 test takers are chosen; the number of candidates for a pre-test is determined by the average number of candidates for the level

in question. Pre-testing is primarily performed at schools and Euroexam test centres. The data provide the basis for the quantitative analysis, which includes statistics for total scores and item-level statistics (**Step 11**):

1. facility values
2. discrimination
3. reliability
4. distractor analysis
5. standard deviation
6. SEM

The facility value or p-value of an item is expressed as a percentage of candidates who answered the item correctly. The p-values of the item are expressed in a range of 0-100%.

The discrimination index gives information about the extent to which the item discriminates between candidates of different ability. The value ranges between -1 and +1, and it gives the degree of association between the candidates' scores on the item and their scores on the test as a whole.

Reliability is the measure of the test items' internal consistency. The estimate is called Cronbach's alpha and ranges between -1 and +1. Ideally reliability should be around 0.9, but the test papers are acceptable above 0.75.

The distractor analysis provides the proportion of test takers who selected each of the response options of a multiple-choice test. Distractors function appropriately if they effectively draw away test takers from the correct answer. Basic evaluation criteria are as follows:

- The key is the most frequently chosen response.
- All response options are functional, i.e. each response is chosen by at least one test taker.
- Item facility values range between 0.90 and 0.10, otherwise they are regarded as extreme and the items are potentially excluded from further analyses.
- Item quality as reflected by the item-test correlation exceeds 0.20.
- Item quality as reflected by the item-rest correlation is positive.
- Distractor quality as reflected by the distractor-rest correlation never exceeds 0.10.
- The item-rest correlation is higher for the key than all distractor-rest correlation coefficients.

Descriptive statistics are calculated for total scores. The most important values are (a) the

mean, (b) the standard error of the mean, (c) the standard deviation, and (d) the reliability estimate with its associated standard error.

The pre-testing documentation also includes a student questionnaire (Appendix 3), which has several purposes. On the one hand, Euroexam International collects additional data on candidate language learning background, age and schooling; on the other hand, students are invited to comment on the pre-test version of the exam.

As part of the student questionnaire, participants confirm that they will not register for a Euroexam test at the pre-test level within the following six months. After the registration deadline for the live test in question, names are cross-checked to ensure that no such registration is made. If this happens, the candidate is offered an alternative testing date.

The results of candidates on pre-tests are produced using the same method as the live examination and sent to the schools for their information.

2.3.4 Second moderation

The results from pre-test statistical analysis provide the basis for the second round of moderation where items with undesirable properties are excluded from further use.

When the results from a pre-test have been analysed, panels of item reviewers consider the item statistics and decide on the final items /tasks to be included in the question paper (**Step 12**). On the basis of the empirical data, items/tasks are accepted, referred for amendment or rejected.

Reviewers first consider the facility value (p-value) for each item. If the p-value is unexpectedly low, the item is reviewed to see if its presentation is confusing or misleading, or if another option could also possibly be correct. In the case of multiple-choice items, if the p-value is unexpectedly high the item is reviewed to see if the stem of the item includes a clue to the key, or if the distractors are so implausible that they can be discounted easily. Items with a p-value below 0.10 or above 0.90 are not accepted.

Reviewers also consider the discrimination index. For testing purposes, items closer to 0.5 are considered highly appropriate for item selection. A discrimination value greater than 0.2 implies the question is performing well, whereas closer to zero, or lower than zero, implies it is not discriminating sufficiently between candidates of different ability. A negative value indicates that the candidates with poorer performance on the rest of the test are performing better than those with a better overall performance on the test. In this case the item is

reviewed to see if its presentation is confusing or misleading. When interpreting the value of discrimination, it is important to be aware that there is a relationship between an item's difficulty index and its discrimination index. If an item has a very high (or very low) p-value, the potential value of the discrimination index will be much less than if the item has a mid-range p-value. Items with a discrimination index below 0.2 are not considered acceptable.

The panel of the second moderation also checks the graphics and the recording of the listening script and may ask the illustrator or the studio for amendments (**Step 13-14**).

2.3.5 Finalisation of test papers

The finalisation of test papers (**Step 15**) involves final editing and proofreading by the Test Production Coordinator and a final content check by the Test Production Manager. A third member of the assessment team (ideally a person who has not been involved in the process) proofreads the finalised material and ensures that no typos or mistakes are left in the papers. The finalised papers are stored on Euroexam secure LAN servers.

Another crucial element of test assembly is the inclusion of anchor tasks. These tasks enable the Assessment Team to gain more information about the test and the test taker, and they are also an indispensable element of statistical analysis for grade reporting.

3. Test Delivery

Euroexam International takes exam security very seriously for the benefit of its customers and the educational institutions and employers who might require proof of language proficiency. All persons involved in the development, administration and execution of the examination must sign a confidentiality agreement. All materials are saved on a secured server, which is only accessible to authorised staff members. Candidate data are stored in a custom-made management tool (LEO), which is only accessible to authorised staff.

The delivery of the exam (**Step 16**) is governed by a detailed set of regulations from start to finish. The *Euroexam Handbook and Code of Practice* and the processes described in the Exam Delivery System cover the following points:

1. Arranging venues
2. Registering test takers (IT facilities)
3. Sending materials (document trails, access control, handling physical test materials, anti fraud processes)
4. Administering the test
5. Returning test materials

4. Marking

In order to maintain the highest possible level of professionalism in the assessment of test takers' written work, all papers are returned to the Euroexam Test Production Centre and are randomly allocated to markers. The aim of marking (**Step 17**) is to assess test taker responses and performance and provide reliable and unbiased results.

The MCQs of the Reading and Listening papers are marked by Optical Mark Recognition and short answers are marked by human assessors. Both objective and subjective marking is done by trained professionals. Subjectively marked tasks are assessed by two standardised markers, who score the papers separately (blind double marking).

The details of the marking process are described in the *Marker standards*. The document provides details on the following:

1. Marker recruitement (necessary qualifications, experience)
2. Marker training (schedule and material)
3. Marker standardisation process
4. Managing the marking process

5. Statistical analysis of live tests

Collecting and analysing data from live tests provides the basis of reporting results. (**Step 18**).

In the course of live test data analysis, both Classical Test Theory (CTT - used for all four papers) and Item Response Theory (IRT - used for Listening and Reading) are employed. The CTT analyses used are described under pre-test data analysis and second moderation (**2.3.3.2**; **2.3.4**).

Modern test theory analysis (IRT) is conducted using the OPLM software, which makes the application of the one-parameter model possible even in the case of items with differing discrimination values. The use of repeated tasks ensures that test taker performance and test difficulty are comparable across administrations. The same test structure and task design make the application of the maximum likelihood estimation procedure possible.

First, items with a negative discrimination value are eliminated from further analysis. Items for which previous information is available are also recalibrated as estimation accuracy will increase with a larger dataset. For items reused in two or more administrations, the appropriateness of sampling and the sample effect are verified by comparing the p-values.

In the first round of calibration, former values of the item bank (reference) are compared with the newly computed values as part of the examination of estimation constancy. Then, in case of an acceptable model-data fit, the newly computed item facility and discrimination values are used in OPLM to model the probability of a correct answer on the spectrum of true ability [-1, +1] on test-characteristic curves.

The use of IRT for the analysis of live test results determination has a number of advantages:

1. items are placed on an interval scale measured in logits
2. the difference between items and test takers are measured on a single scale
3. item difficulty can be understood independent of test taker ability

The results of the analysis, together with the use of anchor items help maintain standards in the different exam sessions.

The overall result on the test is the unweighted average of the results on the individual papers. The result is a “pass” if the overall score, that is the arithmetic mean of the four constituent papers, reaches or is higher than 60 points.

6. Item banking

The item bank contains reliable anchor tasks (**Step 19**), which enable the exam development team to gain more information about the test and the test takers, and they are also an indispensable element of the regular posthoc statistical analysis. Euroexam has a sizeable bank of calibrated items (see Appendix 3), and common tasks are selected based on item quality and difficulty. For the sake of test security, no item is reused within a period of 365 days. Statistical test equating ensures that the standard is the same through different testlets.

The steps of the test production and delivery process

		Involved in process												
	Steps	TPC	TPM	IW	MAT	MO	P	ST	VA	IL	PC	TS	TC	MS
1.	1st version of draft item	x	x	x										
2.	1st check (technical+content)		x		x									
3.	Revision of items	x	x	x	x									
4.	2nd check (technical+content)	x	x		x									
5.	1st moderation	x	x		x	x								
6.	Graphics commissioned	x								x				
7.	Recording commissioned	x						x	x					
8.	Graphics added	x												
9.	Recording checked	x												
10.	Piloting, Pre-testing	x					x				x	x		
11.	Statistical analysis of pre-testing results	x	x		x		x							
12.	2nd moderation	x	x		x	x								
13.	(amendments to graphics)	x								x				
14.	(re-recording of listening scripts)	x						x						
15.	Finalisation of question paper	x	x											
16.	Live test											x	x	
17.	Marking												x	x
18.	Live stats, cut score validation, result determination	x	x		x		x							
19.	Item banking	x	x		x		x							

Abbreviations:

IW – Item writer

TPC – Test Production Coordinator

TPM – Test Production Manager

MAT – Member of Assessment Team

MO – Moderators

P – Psychometrician

ST – Studio

VA – Voice actors

IL – Illustrator(s)

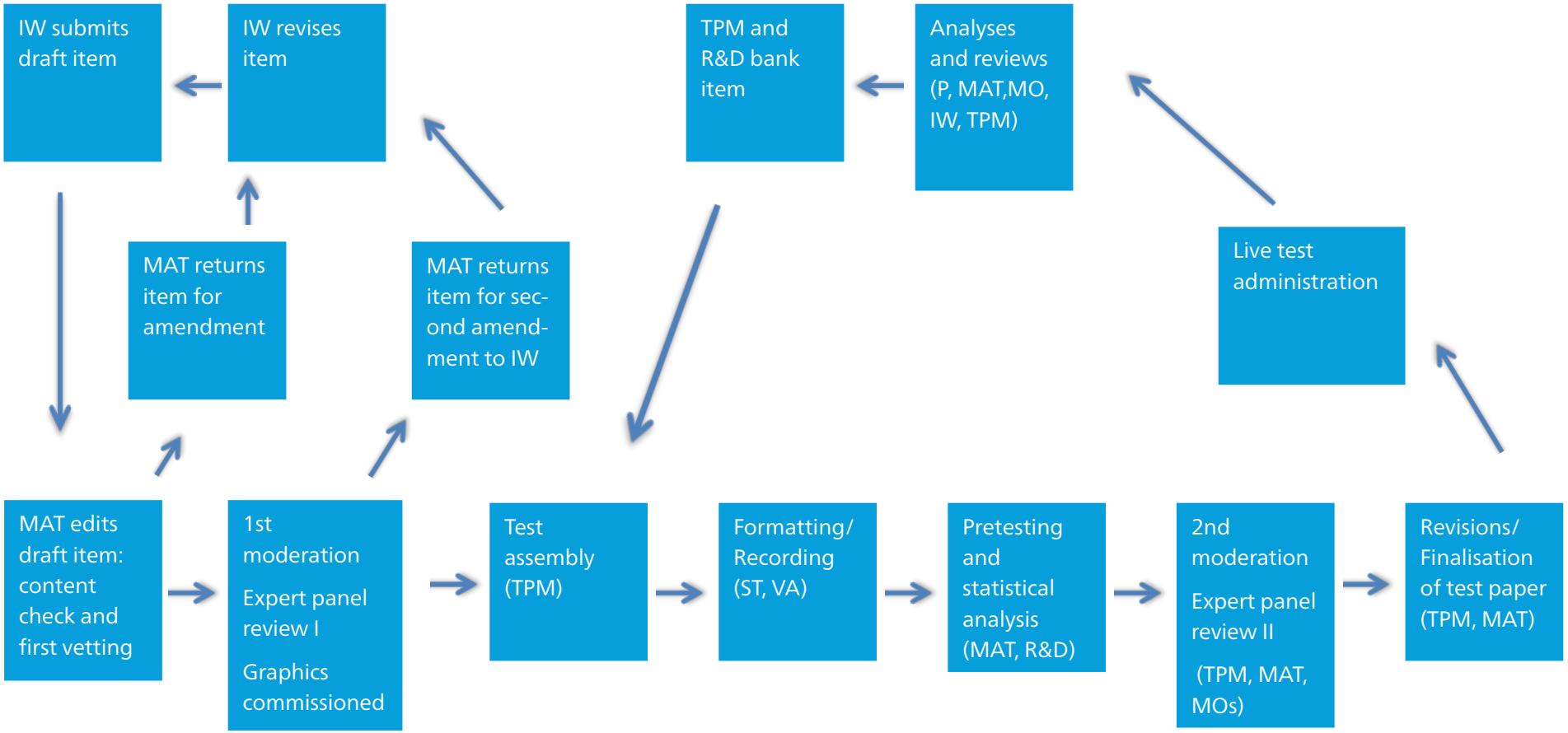
PC – Pre-test Centres

TS – Teachers

TC – Test Centres

MS – Markers

Flowchart of the test production process



Filing system

Filing system and version control

All tasks submitted by item writers for initial vetting are saved to Euroexam secure LAN servers and subsequent versions are stored in an iterative filing system. In conjunction with the Item Writing Schedule, this system ensures that the developmental phase at which any given task is can be easily identified.

Until tasks are accepted to be included in a live paper, their latest versions are stored in the appropriate folder by test type (e.g. general or business and professional), test level and paper (see files in the red frame below). Previous versions are also preserved and stored in the sub-folder entitled *Previous.*

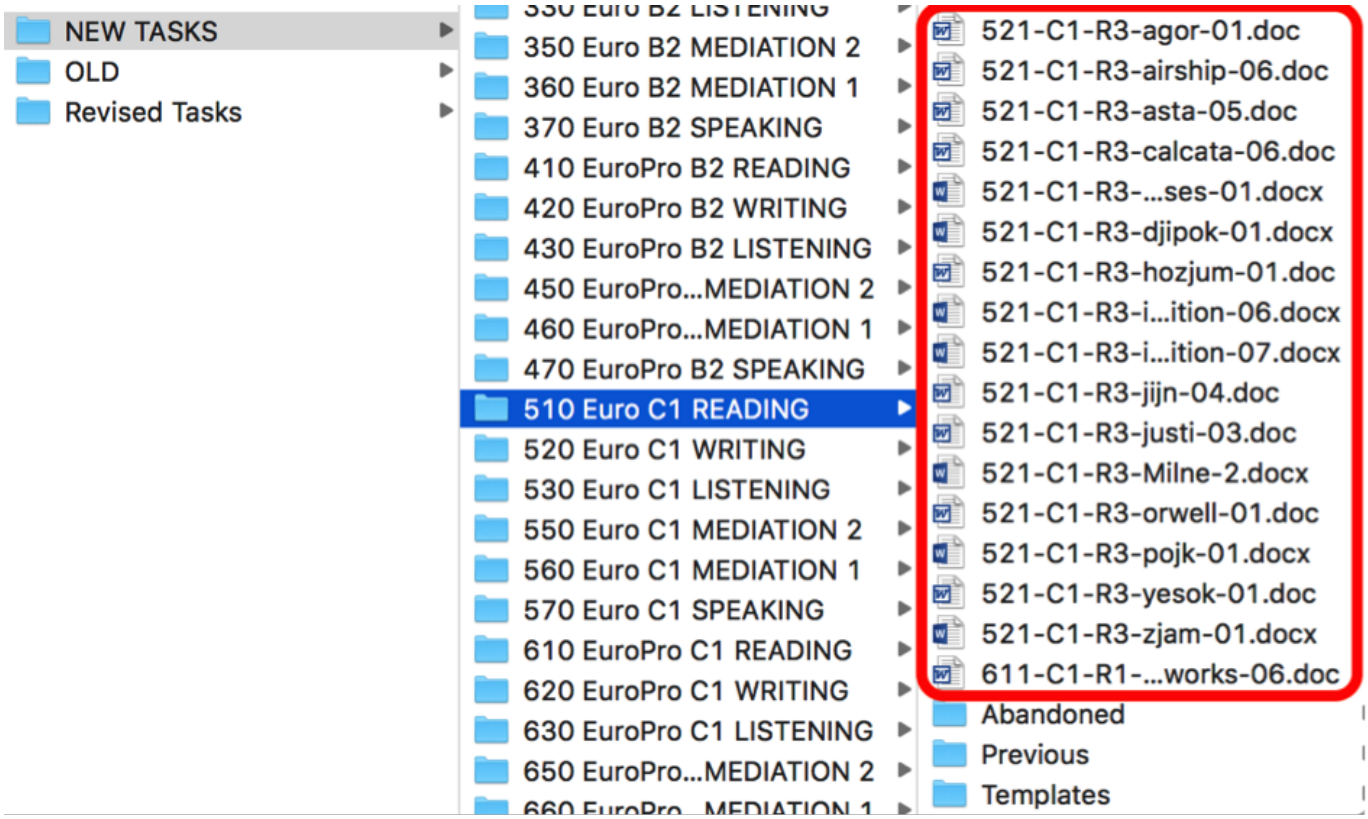


Illustration: the latest versions of Euroexam (General) Level C1 Reading tasks that are at various stages of moderation

The table at the front of all task templates (see below) has clear information about who last changed the task and when the change was made.

513-C1-W1- (Transactional Writing)	
Author	
Task Name *(change as appropriate)*	
Topic of item	
Date created	
Date last modified	
Today's date	
SOURCE (to be attached)	
Document status *(i.e. finished, in progress, or abandoned)*	
Who has checked this item?	
Which exam is this item intended for?	

Key information for writing and checking	
Candidate to write	ca. 200 words
Input pieces	1-2 max 300 words
Time	30 mins

If a task is eventually rejected, it is moved to the folder entitled Abandoned. When a task is selected to be included in a live test, it is moved into a folder entitled *Unformatted*. Then, it is formatted in InDesign and new files are created in the folder entitled *Formatted*.

File naming (Reading, Writing, Listening)

Tasks for the Euroexam Level C1 are written in Word files. Each version of each task has a unique filename, which contains several pieces of information about the task. A typical file name might be:

533-C1-L3-funrip-02

533

The initial 5 shows that this is a C1 task (also indicated by C1 later in the title)

The first 3 indicates that this is a listening task (also indicated by L later in the title)

The second 3 indicates that this is the Radio Programme task.

C1

This shows the level of the task.

L3

This shows that the task is the third task in the Listening Test (namely, Radio Programme)

[N.B. The first three numerals, here 533, are set when the task is designed, and do not change if the order of tasks is subsequently changed. The later designation, here L3, will change if the task order is changed in the exam]

funrip

This is a unique identification key assigned to the task, which expedites searching the Item Writing Schedule and the filing system.

02

This is the version number of the task. Every time substantial amendments are made to the task after the initial draft, either by the original item-writer or someone else, the number is increased by one and the previous version is archived.

Note: Titles of tasks in the Euroexam Level C1 test (Reading, Writing, Listening)

511-C1-R1: Paragraph headings

512-C1-R2: Long text

521-C1-R3: Multiple-choice reading

513-C1-W1: Transactional writing

522-C1-W2: Discursive writing

531-C1-L1: Short conversations

532-C1-L2: Making notes

533-C1-L3: Radio programme

Euroexam International Pre-test Candidate Questionnaire

	Personal data			
Name				
Date of birth (dd/mm/yy)				
Sex (M/F)				
Name of School				
When did you start learning English? (yyyy)				
Indicate the level of your language exam certificate. Circle the level and give the name of the awarding body (e.g. Euro, Origó, etc.)	B1 awarding body:	B2 awarding body:	C1 awarding body:	other level awarding body:

	Your opinion about the pre-test tasks		
Which part of the test did you find most difficult? Why?			
Which part of the test did you find easiest? Why?			
How much time did you need for each part of the test?	Reading ..... min	Listening ..... min	Writing ..... min
Do you think you will achieve 40 or more on each part?			
Do you think you will achieve 60 or more on each part?			

Please rate the test (1= very poor 5= excellent) based on the following:

	1	2	3	4	5
Layout and appearance					
Instructions					
Task types					
Content					

I understand that I cannot take a Euroexam test at this pre-test level within the next six calendar months.

Signed Date