

**Developing English language tests for Luxembourg
secondary schools: The Test Design and Evaluation (TDE)
project, 2011-2014**

Dr Tineke Brunfaut and Dr Luke Harding
Lancaster University, United Kingdom

31 December 2014

Table of Contents

1	Introduction	1
2	Brief history of the project.....	2
2.1	Training	2
2.1.1	Language testing coursework	2
2.1.2	Workshops	4
2.2	Test development	5
2.2.1	Specifications phase.....	5
2.2.2	Item writing, review and moderation phases.....	7
2.2.3	Piloting and pre-test analysis phases.....	7
2.2.4	Operational assessment phase	8
2.2.5	Assessment review phase	8
2.2.6	The overall test cycle of the 2013 <i>Épreuve Commune</i>	9
2.2.7	The 2014 <i>Épreuve Commune</i>	9
3	Quality of the <i>Épreuve Commune</i> for English	11
3.1	The 2013 <i>Épreuve Commune</i> administration.....	11
3.1.1	Listening	11
3.1.2	Reading	13
3.1.3	Teacher feedback.....	14
3.2	The 2014 <i>Épreuve Commune</i> administration.....	14
3.2.1	Overall test.....	14
3.2.2	Teacher feedback.....	15
4	Dissemination	15
4.1	Stakeholder exchange.....	15
4.2	The <i>Épreuve Commune</i> website.....	15
4.3	Cascading language assessment literacy	16
4.4	Communicating beyond Luxembourg.....	17
5	Summary and recommendations.....	17
6	References	19
7	APPENDICES	20
7.1	Appendix 1: 2013 <i>Épreuve Commune</i> teacher questionnaire data	20
7.2	Appendix 2: 2014 <i>Épreuve Commune</i> teacher questionnaire data	23
7.3	Appendix 3: Initial analyses of the 2014 <i>Épreuve Commune</i>	24
7.4	Appendix 4: 2013 English Teachers' Day presentation.....	25

Table of Figures

Figure 1: Test cycle - <i>Épreuve Commune</i> 2013	9
Figure 2: Test cycle - <i>Épreuve Commune</i> 2014	10

Table of Tables

Table 1: 2013 <i>Épreuve Commune</i> listening test item analysis.....	12
Table 2: 2013 <i>Épreuve Commune</i> reading test item analysis	13

1 Introduction

The aim of this report is to describe the activities, achievements and outcomes of an English language secondary school exam reform project in Luxembourg.

In 2010, as part of reforms and development of English language teaching in Luxembourg secondary schools, a number of stakeholders identified the existing system of testing and assessment to be an obstacle in making the desirable progress. In particular, the nature of the *Examen de Fin d' Études* – the end-of-school-leaving exam – was regarded as not well-aligned with the new approaches to English language teaching. At the same time, those involved recognised that there was a relative lack of formal expertise in language testing and assessment amongst the teaching body.

Out of this concern arose the Test Design and Evaluation (TDE) initiative in 2011. A number of teachers on the secondary school curricular board (CNP-ES) proposed to improve the existing end-of-school-leaving exams, whilst at the same time building up language testing expertise within the country as a step towards establishing a more professional basis for the design of tests. To realise these aims, the then Head of CNP-ES, Josiane Weis, approached Lancaster University (UK), and in collaboration with the *Formation pédagogique des enseignants du secondaire* (FOPED) unit at the University of Luxembourg and with the *Service de Coordination de la Recherche et de l'Innovation pédagogiques et technologiques* (SCRIPT) at the *Ministère de l'éducation nationale et de la formation professionnelle*, a proposal was formulated for a joint project which led to the establishment of the TDE group. At present, the team consists of representatives from SCRIPT (Maurice Broers, Thomas Michels, and Romy Schmit), FOPED (Michel Fandel and Halldor Halldorsson), and from various English language school streams (Michel Dohn, Christiane Klein, Phillippa Probyn, Marc Trierweiler, and Anne Zimmer). As such, three key types of stakeholders take an active part in shaping and conducting the TDE project. The authors of this report, Dr Tineke Brunfaut and Dr Luke Harding, have provided consultancy and training throughout, and can therefore be considered adjunct members of the team.

In this report, we will first describe the activities that have taken place during the first three years of the project, from September 2011 until August 2014. The nature and extent of the training conducted with the TDE team will be specified in section 2.1, and the development work leading to the new *Épreuve Commune* for English will be detailed in section 2.2. Section 3 reviews the first version of the *Épreuve Commune* administered in June 2013, and comments on the overall quality of the test. The team's work on disseminating the knowledge and skills they gained, as well as the instruments, procedures and tools they developed is described in section 4. In the final part of this report (section 5), a number of recommendations are made to ensure the sustainability of the high-quality assessment work that has been accomplished so far. The TDE skills basis to progress to a high-stakes test context is also commented upon.

2 Brief history of the project

From the start, the TDE initiative's central and long-term aim has been to develop an end-of-secondary-school-leaving exam that aligns with and forms a core part of the national curriculum reforms for English – reflecting current approaches to language teaching and learning, including an orientation to the Common European Framework of Reference (CEFR; Council of Europe, 2001). At the same time, given the stakes of the *Examen de Fin d' Études* for individual language learners as well as for the Luxembourg educational system's international reputation, the TDE initiative has from the outset emphasised the need for a high-quality test which can live up to the criteria of good practice as defined by international bodies such as the Council of Europe, the European Association of Language Testing and Assessment (EALTA) and the International Language Testing Association (ILTA).

In consultation with language testing experts at Lancaster University, and in order to realize this ambitious undertaking, it was considered vital to establish a solid knowledge basis and to develop the know-how of those involved regarding the production of high-quality, professional language tests. Therefore, a language testing training programme was set up to develop the language testing expertise of the TDE team members. The nature of this training programme is described in section 2.1 below. In addition, given the high stakes of the *Examen de Fin d' Études* it was considered important that the team would first gain experience in setting up a test cycle and developing a lower-stakes test. Section 2.2 describes the activities that have been undertaken to develop the lower-stakes *Épreuve Commune* and to design and implement a test cycle for rolling out the exam nationally.

2.1 Training

The development of the TDE team's language testing knowledge and skills was led by Dr Tineke Brunfaut and Dr Luke Harding from the Department of Linguistics and English Language at Lancaster University, who are internationally known for their expertise in language testing training and research. The TDE training took two different forms. On the one hand, the team members took three 9-unit courses to develop their theoretical and practical knowledge in language testing (2.1.1). On the other hand, they participated in three hands-on workshops on aspects of a professional test cycle (2.1.2).

2.1.1 Language testing coursework

Three courses covering a wide range of theoretical and practical aspects of language testing were taken by the members of the TDE team. Each of the courses was equivalent to a 20-credit Masters level course in the UK Higher Education context. In each school year of the period 2011-2014, the team took one course. The knowledge and skills focussed on in each course were aligned with the on-going development activities of the *Épreuve Commune*. The courses were:

- Language Test Construction and Evaluation (October 2011 – February 2012)
- Issues in Language Testing (September 2012 – January 2013)
- Researching Language Testing (January 2014 – May 2014)

The first course, *Language Test Construction and Evaluation*, provided a solid introduction to language testing. The aims of this course were to familiarise the team with the ways in which tests are designed and how test data is analysed, to enable them to construct and evaluate their own tests, and to enable them to critically evaluate test items. The course introduced fundamental concepts and terminology in language testing, and acquainted the team with the various phases of the test design process. In addition, the team learned how to qualitatively analyse tests and items, and how to run and interpret descriptive statistics on test and item performances. They learned how to combine these various pieces of information and insights to evaluate the quality of tests with the aim of improving them. Furthermore, specific attention was given to considerations and good practice in the testing of reading, writing, listening and speaking skills.

The course was partly delivered face-to-face in Luxembourg, and partly through Lancaster University's online learning platform. The course was divided into 9 learning units, each consisting of compulsory academic readings, reflective and discussion tasks, and hands-on development and evaluation tasks which needed to be posted to the online platform by the end date of the unit. Feedback on the work was provided by the course instructors.

Dr Brunfaut visited the team in Luxembourg at the start of the project in October 2011 to familiarise the team with the online learning platform, the statistical software, and to deliver the first unit of the course. Dr Harding conducted the final unit's training face-to-face in December 2011. To evaluate the team's overall learning on the course, each person individually conducted a test evaluation study in January 2012, which was assessed by the course instructors.

The second course, *Issues in Language Testing*, extended the team's theoretical and practical language testing knowledge and skills by focusing on a selection of more advanced topics. This course aimed to further consolidate the team's understanding and application of key concepts in language testing, to make them familiar with various ways of testing and assessing language ability, and to familiarize them with different approaches to validity. The course was delivered through Lancaster University's online learning platform, following a similar 9-unit structure as the first course and consisting of readings and tasks on which feedback was given by the course instructors. The course content included a review of language testing concepts, the writing of constructed- and selected-response items, the testing of grammar and vocabulary, and alternative ways of assessing language ability. Importantly, the team learned about traditional and more recent approaches to validity and validation, and standard setting theory.

Having gained a considerable amount of knowledge and skills in language testing, the third course – *Researching Language Testing* – aimed to equip the team with the knowledge and skills necessary to undertake future research on the tests they develop, and to support the team members in becoming independent language test evaluators and researchers. The course was delivered through Lancaster University's online learning platform, again structured along 9 learning units and consisting of readings and tasks on which feedback was given by the course instructors. The course dealt with the theoretical, methodological and ethical issues that are central to research in language teaching and language testing. The team was familiarized with and tried out various ways of collecting and analysing qualitative and quantitative data resulting from research on language tests. This included questionnaire, interview and verbal protocol methods, as well as more advanced statistics to look into test and rater reliability and comparative statistical techniques. Being

able to investigate the quality of the tests one develops is vital to establishing their validity and ensuring that valid and fair conclusions are drawn on test-takers' language ability on the basis of their test performances.

2.1.2 Workshops

In addition to the courses, Drs Brunfaut and Harding offered three workshops to the TDE team members on two critical stages in any test cycle (see section 2.2 below): item moderation and standard setting. The aim of these workshops was to give the team first-hand experience and insights into the focus, goal, and methodology of these test cycle stages, and at the same time to give the team a chance to observe how these are professionally run. The ultimate goal was to equip the team with the knowledge and skills to organize and lead these test cycle stages without external help. Thus, the workshops aimed to enable the longer-term sustainability of the TDE endeavour. At the same time, the item moderation and standard setting workshops were organised when these test cycle stages were reached in the development cycle of the *Épreuve Commune* and worked with the *Épreuve Commune* materials. In this manner, the work undertaken during the workshops constituted the execution of key components of the test cycle.

All workshops were conducted face-to-face in Luxembourg and convened by the Lancaster University experts. The item moderation workshop took place on 11th and 12th January 2013. Item moderation has as its purpose to review the quality of language test items that have been developed on the basis of a set of test specifications and to evaluate the correspondence of the items with the specifications. As Green (2014: 43) emphasizes: "All material should pass through at least one round of item review (...) before it is used in earnest." During the workshop, the reading and listening items the TDE members had written in two groups (each group specialising in one skill) were scrutinised by the other group. During the first two courses the team members had taken (see 2.1.1), they had gained knowledge and skills for qualitative item evaluation. At the workshop, the team assessed the items' quality, checked the items against the specifications and the CEFR, identified their strengths and weaknesses in relation to characteristics of good-quality items, and made suggestions for improvement where relevant. This allowed the team to then progress to the next stage of item revision and task selection for trialling (see 2.2). The workshop was concluded with a discussion of the item moderation experience and an opportunity for clarifying questions on the methodology.

The second and third workshops focussed on standard setting. Theoretical knowledge on standard setting and some initial insights into procedures had been acquired as part of the *Issues in Language Testing* course. The first workshop was held on 13th and 14th September 2013, after the first administration of the *Épreuve Commune* in June 2013 and statistical analyses had been conducted on the results. The second workshop took place on 9th and 10th May 2014, working with the pre-tested items and statistical results, in order to compile the test version for the second administration of the *Épreuve Commune* in June 2014. Apart from the TDE team, external stakeholders had also been invited to participate in the standard setting workshops, since the representation of a variety of stakeholders and the combination of insider and outsider perspectives is an important feature of standard setting (Brunfaut & Harding, 2014). The then president of the vocational secondary school curricular board, the CNP-EST, participated in the 2013 item moderation and standard setting workshops, and the president of the CNP-ES took part in the 2013 standard

setting. In addition, a vocational school English language teacher joined the item moderation workshop and two more English language teachers attended the 2014 standard setting.

The aim of a standard setting procedure is to establish a performance standard and determine one or more cut scores through expert judgements on the level of the items. The specific procedure adopted was that of the Modified Basket Method, which has been shown to be practical, reliable and valid for the standard setting of receptive skills items (Brunfaut and Harding, 2014). Level judgements were made with reference to the CEFR (the framework underlying the test's development), and workshop participants were familiarised with the CEFR and the standard setting procedures prior to the actual standard setting. The workshops prioritised the Familiarisation and the Standardisation stages of the Council of Europe's (2009) manual for relating exams to the CEFR. The practical outcome of the second workshop was a validation of the level and cut scores applied to the 2013 *Épreuve Commune*. The third workshop resulted in standard set items and tasks that informed the compilation of the 2014 *Épreuve Commune* and its cut score(s). Similar to the first workshop, both standard setting workshops were concluded with a discussion of the standard setting experience and an opportunity for clarifying questions on the methodology.

2.2 Test development

The development of a useful language test (as defined by Bachman and Palmer, 1996) is a lengthy process with several checks and balances. Green (2014: 42) describes a typical assessment production process as a cycle consisting of a minimum of seven phases:

- 1) a specifications phase,
- 2) an item writing phase,
- 3) an item review phase,
- 4) a piloting phase,
- 5) a pilot review phase,
- 6) an operational assessment phase,
- 7) and an assessment review phase, which in turn feeds back into the specifications (phase 1).

As explained in section 2 above, in consultation with the experts, the TDE team decided to develop a rigorous test cycle for the lower-stakes *Épreuve Commune* in the first instance. This enabled the team to gain experience in setting up a test cycle and in large-scale test development and administration, before moving on to the high-stakes context of the *Examen de Fin d'Études*.

2.2.1 Specifications phase

Work on the *Épreuve Commune* took off in 2012, after the team members had acquired a theoretical and practical basis in language testing through the first course (see 2.1.1). In the first place, this involved determining within the team and in consultation with various stakeholders what the role, stakes and overall focus would be of the English *Épreuve Commune*. The decision was made to initially restrict the exam to the testing of learners' English reading, listening, and writing skills. Once the general direction and scope of the exam had been outlined through a process of discussion, reflection and consultation, the team began to set up a test cycle.

Following guidelines of good practice, the first step concerned the development of test specifications for the *Épreuve Commune*. The initial efforts concentrated on specifications for the

reading and listening sections of the exam. The specifications for the writing section were compiled at a later stage (2012-2013), after the team had tried out developing reading and listening items on the basis of the specs and had had an opportunity to evaluate the functionality of the corresponding specification documents.

The test specifications framework as defined by Alderson, Clapham and Wall (1995) was adopted for this purpose. This involved a decision-making process encompassing a wide range of aspects of the exam and defining the following characteristics:

- The purpose
- The construct
- The target population
- The target level
- The skills to be tested
- Time for each section
- For listening/reading:
 - The nature of input texts
 - The text topics
 - The text length
 - Speaker characteristics (listening)
- For writing:
 - The nature of task prompts
- The number of tasks and items
- The test methods
- The weighting of items
- The instructions
- The response format
- The marking:
 - The criteria
 - The scale (writing)
 - The rating procedures
 - The raters – selection, training, standardisation
 - Reporting
- Test administration details:
 - The physical conditions
 - Test preparation
 - Uniformity of administration
- Sample papers
- Samples of students' performance on tasks

The considerations and decisions made as part of this phase were guided by Bachman and Palmer's (1996) five key principles of test usefulness – validity, reliability, practicality, authenticity, and potential for positive washback – and aimed at striking a good balance between these principles. The process involved a considerable amount of brainstorming, drafting and revision. At several points, the external consultants, Drs Brunfaut and Harding, provided feedback and formulated suggestions on the specification drafts. The resulting specifications (version dated 19/04/2013) have been made publicly available and can be consulted by stakeholders (including teachers, parents, and learners) on: <https://portal.education.lu/Portals/22/English/Documents/20130419%20General%20Specifications.pdf>.

2.2.2 Item writing, review and moderation phases

On the basis of the first comprehensive set of specifications for the reading and listening sections of the *Épreuve Commune*, the TDE team's test development activities in the second half of 2012 primarily consisted of writing, reading and listening tasks and items. The team was split into two groups, each specialising in one of the two skills. The task development also involved devising a professional layout for the tasks and test booklets, and recording procedures for the listening section.

The initial collection of tasks was externally reviewed by the consultants and by Lancaster University's Language Testing Research Group (<http://wp.lancs.ac.uk/ltrg/>). Suggestions for revisions were formulated, which informed redrafting of tasks. On the basis of this first item writing experience, additional tasks were developed, resulting in twice the number of tasks and items that would be needed in the final version of the test. As explained above (2.1.2), all items were then moderated during a workshop held in January 2013, and revised where needed.

Following the knowledge and skills gained in writing the reading and listening items, the team developed writing tasks in line with the test specifications. Feedback was given on the draft tasks by the external consultants, and the tasks were adapted accordingly.

A major component of the writing section development concerned the drafting of a rating scale, including deciding on the number and kind of criteria and defining level descriptors. The drafting process was guided by the principles of scoring validity, reliability, and practicality. TDE members took the lead on the development of the scale, with comments and suggestions being provided by the external consultants at various points. The resulting marking grids can be consulted on: <http://portal.education.lu/LinkClick.aspx?fileticket=KPiscpusafQ%3d&portalid=22>.

2.2.3 Piloting and pre-test analysis phases

The listening and reading tasks were compiled to form two versions of the test, in accordance with the specifications. The booklets were professionally edited, printed and distributed by the team. The two test versions were piloted on 361 and 282 learners respectively from different schools and school systems. The pilot tests were marked by the TDE team, and data were entered into spreadsheets. Descriptive test and item statistics were run on the data by the Lancaster University consultants, and the findings were interpreted according to conventional guidelines (see e.g. Alderson, Clapham, & Wall, 1995; Green, 2013).

On the basis of the pilot study results, the best performing tasks and items were selected, and minor modifications were made to the tasks, in order to compile the final version of the 2013 *Épreuve Commune*. The booklet and listening recordings were prepared, duplicated, and distributed to all participating schools.

In addition, the team developed an assessment booklet and answer form for teachers, explaining the aim and nature of the exam. This booklet also included the answer keys and rating scale, and an explanation of how to calculate the test results. Furthermore, sample writing performances were given with the corresponding completed writing grids and an explanation of the rating decisions. The 2013 assessment booklet and form are available at:

<https://portal.education.lu/Portals/22/English/Documents/Assessment%20booklet.pdf> and <https://portal.education.lu/Portals/22/English/Documents/Printable%20assessment%20sheet.pdf>.

Importantly, the team prepared a feedback booklet for teachers of participating classes. The booklet consisted of a total of 68 five-point Likert scale questions asking for teachers' views on the listening, reading, and writing sections of the test; on overall test; on the test administration; and on the marking keys, grid, and concurrent validity. In addition, participating teachers were given the opportunity to add further comments in an open-ended section at the end of the questionnaire.

In 2014, the writing scale was externally validated by the Language Testing Research Group at Lancaster University by applying the scale to a set of written performances from the June 2013 *Épreuve Commune* administration.

2.2.4 Operational assessment phase

The 2013 *Épreuve Commune* test booklet and listening recordings have been made available post-hoc on the following website:

<http://portal.education.lu/epreuvescommunes/English.aspx#3294607-epreuve-commune-20122013>. The test and assessment booklets were sent out to schools who volunteered to take part in the centrally developed exam. 1215 learners (355 from the *Enseignement Secondaire* (ES) system and 860 from the *Enseignement Secondaire Technique* (EST) system) completed the exam in June 2013. 18 teachers returned their completed feedback questionnaires.

The marking was conducted by the learners' teachers, and the teachers entered the item results into a national database system that is conventionally used for this purpose. Scores and completed booklets were archived by the *Ministère de l'éducation nationale et de la formation professionnelle*.

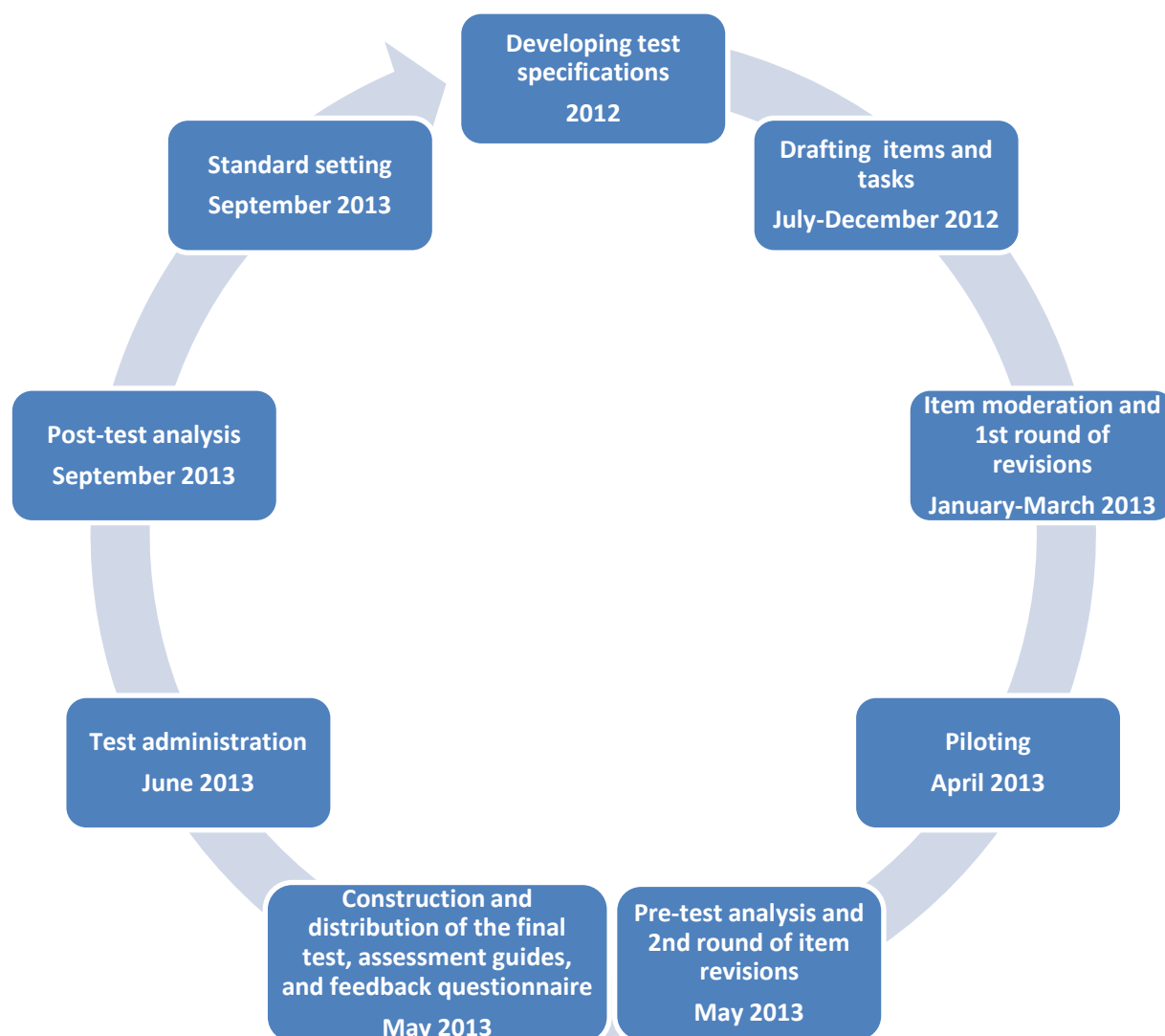
2.2.5 Assessment review phase

Post-test analysis was conducted in September 2013. Statistical test and item analyses were run by a TDE team member who is employed as a psychometrician at the *Ministère de l'éducation nationale et de la formation professionnelle*. In addition, the teacher feedback questionnaires were analysed. The item statistics formed part of the materials that fed into the standard setting workshop held that same month (see 2.1.2). The test results were also more generally interpreted by Drs Brunfaut and Harding, and presented as part of an overall team discussion. A summary of the assessment review analyses is provided in section 3 below.

2.2.6 The overall test cycle of the 2013 *Épreuve Commune*

The test cycle followed for the 2013 *Épreuve Commune* can be summarised as shown in Figure 1.

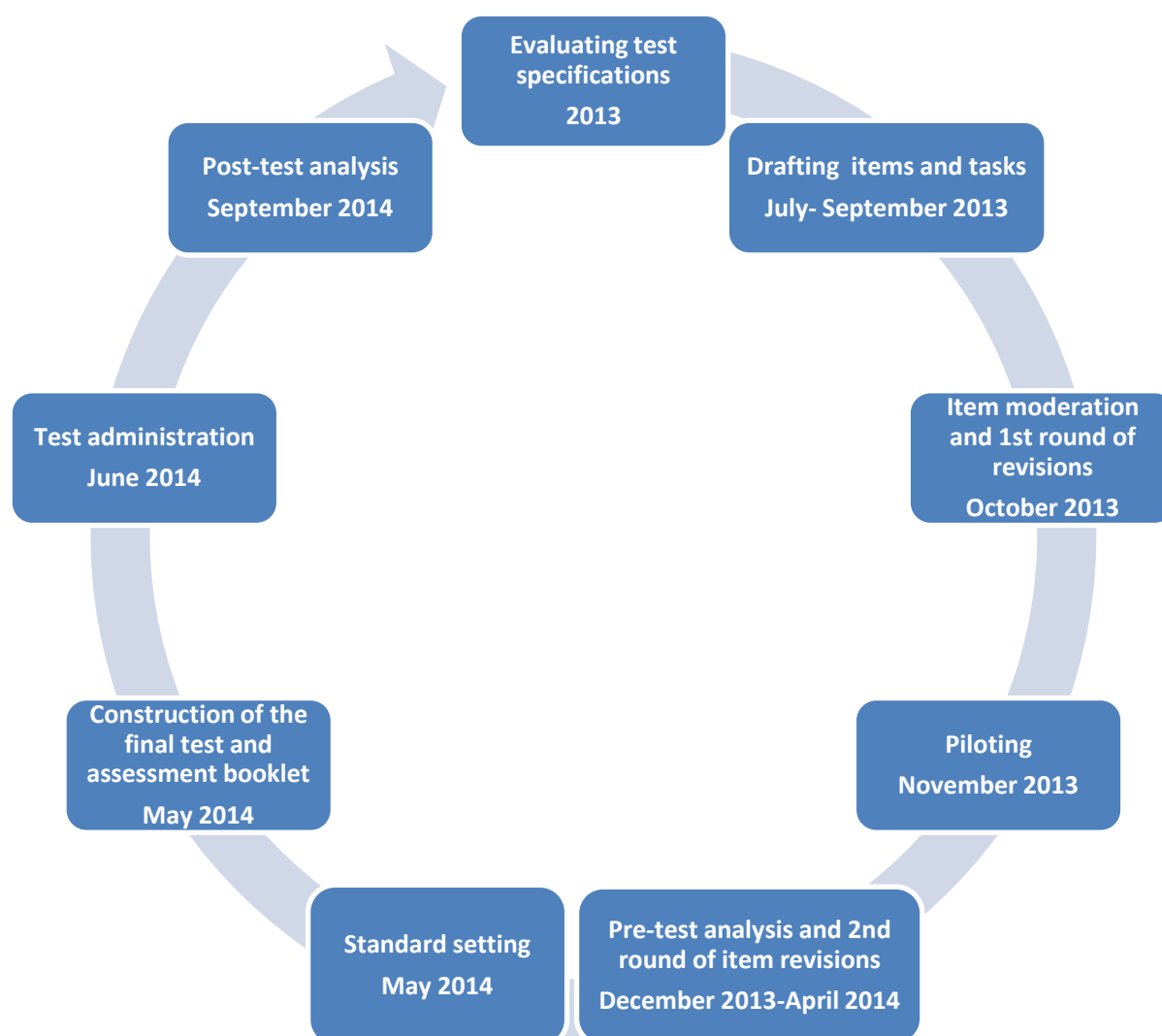
Figure 1: Test cycle - *Épreuve Commune* 2013



2.2.7 The 2014 *Épreuve Commune*

On the basis of having developed and conducted the first full test cycle in 2012-2013, and the successful experiences with the 2013 *Épreuve Commune*, the team predominantly carried the activities of the 2013 test cycle across to the development cycle of the 2014 *Épreuve Commune*, with minor changes in the exact timing of each activity. One key difference was the timing of the standard setting, i.e. after the piloting stage (which is in fact a more conventional sequence of phases). In this manner, the standard setting could inform the selection of tasks for the final exam and the cut score decisions. It is expected that this test cycle will guide the development of future editions of the *Épreuve Commune*.

Figure 2: Test cycle - *Épreuve Commune* 2014



Similar to the 2012-13 development, the majority of the work on the 2014 *Épreuve Commune* was carried out by the TDE team, in consultation with the Lancaster University consultants, Dr Brunfaut and Dr Harding. Two versions of the tests were piloted on a total of 350 learners, representing 5^{ème} ES classique (115), 4^{ème} ES modern (52), 10^{ème} EST Régime technique (148), 10^{ème} EST Régime de la formation du technicien (35). The learners who were selected for the piloting were slightly further or earlier in their schooling than the test target population to avoid that the target population would be familiar with the test tasks prior to official *Épreuve Commune* administration. The final version of the *Épreuve Commune* was administered in June 2014 to 1828 learners, i.e. 1222 learners from the various classes of the EST system and 606 from the ES system.

The 2014 *Épreuve Commune* materials have been made publicly available at: <http://portal.education.lu/epreuvescommunes/English.aspx#3294541-epreuve-commune-20132014>. This includes the test booklets, assessment booklets and grids, and listening test materials.

An important addition to the materials constitutes the training materials for teachers on using the marking grids to rate the written performances:

<https://portal.education.lu/Portals/22/English/Documents/TRAINING%20MATERIALS.pdf>.

3 Quality of the *Épreuve Commune* for English

As well as ensuring procedural quality in the test development process (as described in the foregoing sections), test quality can also be checked post-hoc in several different ways: through statistical analyses of reliability, through item analysis, and through stakeholder feedback. Often various pieces of evidence are combined to arrive at a judgement as to overall quality and areas where improvements may be made. In this section we will discuss the evidence yielded by the 2013 administration of the *Épreuve Commune* with a view to establishing its empirical quality as an instrument that is fit for purpose.

3.1 The 2013 *Épreuve Commune* administration

The first *Épreuve Commune* was administered in 2013 to 1171 students. The results for the listening and reading paper, along with a brief description, are provided below.

3.1.1 Listening

According to the specifications developed in the process described above, the listening test consists of four tasks. In Task 1, test-takers hear a series of short monologues or dialogues on familiar, everyday topics and need to complete a multiple-choice picture selection task for each passage. The focus of the individual questions (henceforth “items”) is on understanding details in the text, and recognising word- or phoneme-level information. In Task 2, the test-takers again listen to short monologues or dialogues and answer multiple choice questions or complete matching tasks. In this task the focus is on understanding the main ideas of a passage, or the “gist” (overall meaning). In Task 3, test-takers hear a longer monologue and answer a range of comprehension questions. For the 2013 *Épreuve Commune*, the test-taker responded to this task by selecting True, False or Doesn’t Say for a range of statements. Task 4 is a longer dialogue, and test-takers respond by completing a form with the information they have heard. Overall, the test takes approximately 30 minutes to complete (including instructions and pauses, which are all pre-recorded on the CD).

The Listening section of the 2013 *Épreuve Commune* contained 28 items and had an overall reliability of .74 (Cronbach’s alpha). This reliability figure is a measure of the internal consistency of the test; the extent to which the items on the test measure the same construct (in this case listening). Reliability ranges from 0 (no reliability) to 1 (perfect reliability). A figure of .74 is considered satisfactory for a test which is being used for the purpose for which the *Épreuve Commune* is designed, that is, for pedagogical purposes, and not to make major decisions regarding progression (where figures higher than .8 would be desirable). It is also worth noting that, as the TDE team’s first attempt at designing a large-scale test, this is a very respectable figure indeed.

Table 1 shows the facility values and item-total correlation for each of the items on the listening test. The facility value (FV) is a measure of item difficulty ranging from 0 (an item that no-one got correct) to 1 (an item that all test-takers got correct). Therefore, the FV represents the proportion of test-takers who got a given item correct, which is taken as an indication of the

difficulty of that item. The item-total correlation is a measure of item discrimination: the degree to which an item distinguishes between high-ability and low-ability test-takers. The item-total correlation ranges from -1 to 1. A value of 1 would indicate that high-ability and low-ability candidates are being distinguished perfectly by the item; a value of 0 would mean that the item is not distinguishing between ability levels at all; and a value of -1 would mean that low-ability candidates were getting the item correct while high-ability candidates were getting it wrong (which would indicate a major problem with the item). Together, these statistics which gauge difficulty and discrimination are fundamental indicators of item (and therefore test) quality.

Table 1: 2013 Épreuve Commune listening test item analysis

Item	FV	Item-total correlation
LT1_1	0.86	0.20
LT1_2	0.86	0.29
LT1_3	0.68	0.36
LT1_4	0.68	0.20
LT1_5	0.99	0.17
LT1_6	2.60	0.20
LT2_1	0.98	0.26
LT2_2	0.98	0.28
LT2_3	0.88	0.25
LT2_4	0.98	0.30
LT3_1	0.90	0.19
LT3_2	0.85	0.24
LT3_3	0.84	0.36
LT3_4	0.65	0.22
LT3_5	0.82	0.12
LT3_6	0.69	0.21
LT3_7	0.77	0.25
LT3_8	0.66	0.34
LT4_1	0.54	0.30
LT4_2	0.93	0.32
LT4_3	0.32	0.24
LT4_4	0.73	0.41
LT4_5	0.58	0.32
LT4_6	0.50	0.31
LT4_7	0.85	0.24
LT4_8	0.88	0.14
LT4_9	0.99	0.11
LT4_10	0.94	0.21

According to these results, the test items were generally quite easy across the board for the population. 16 out of the 28 items had FVs above .80, and the average FV was .79¹, which suggests the test-takers had mastered the level of this test – which was pitched at A2 level on the CEFR – well. This high clustering of FVs would also have an impact on the reliability of the test, as it is difficult to

¹ Not including item LT1_6, which was scored as partial credit and therefore had an FV over 1.

achieve a very high reliability without a spread of scores. The discrimination values show that the items were of a generally good quality. 23 out of the 28 items had item-total correlations over .20, which suggests that the majority of items were discriminating sufficiently well for a test of this kind. Importantly, there were no negatively discriminating items which suggests (a) that the items were well-designed to tap into the listening construct and were not testing other skills or abilities, and/or (b) that the items were free of serious flaws such as more than one correct answer in a multiple-choice question.

3.1.2 Reading

Like listening, the *Épreuve Commune* reading test consists of four tasks. In Task 1, test-takers read short texts for general understanding and gist. In Task 2, test-takers show they have understood signs, notices and other short texts which might be found in the public domain, demonstrating understanding of specific information in these texts. Tasks 3 and 4 have the same broad specifications: test-takers read longer texts for comprehension of details and main ideas, and respond on multiple choice questions or by selecting True, False or Doesn't Say. Texts, again, are on familiar topics, with the 2013 *Épreuve Commune* drawing on an advertisement recruiting teenage fashion models, and an email from a student on exchange in London.

The analysis of data from the live administration showed that the reading section of the 2013 *Épreuve Commune* was, on the whole, a good, solid reading assessment (see Table 2). The overall reliability of the section was .81 (Cronbach's alpha), which is a very solid figure for a test of this nature, comparable to many other large-scale tests. As with the listening test, the items were generally easy across the board, although the average FV was somewhat lower than the listening test at .76.

Table 2: 2013 *Épreuve Commune* reading test item analysis

Item	FV	Item-total correlation
R1_1	0.89	0.43
R1_2	0.85	0.50
R1_3	0.86	0.49
R1_4	0.84	0.50
R1_5	0.81	0.56
R1_6	0.63	0.45
R1_7	0.90	0.48
R2_1	0.66	0.43
R2_2	0.61	0.30
R2_3	0.86	0.15
R2_4	0.93	0.40
R2_5	0.91	0.44
R2_6	0.53	0.24
R2_7	0.88	0.32
R3_1	0.86	0.27
R3_2	0.56	0.43
R3_3	0.71	0.24
R3_6	0.61	0.27
R3_7	0.88	0.38

R3_8	0.85	0.36
R4_1	0.80	0.26
R4_2	0.94	0.25
R4_3	0.59	0.29
R4_4	0.94	0.22
R4_5	0.77	0.25
R4_6	0.52	0.20
R4_7	0.71	0.24
R4_8	0.44	0.21

As with the listening test, there were few items which had an item-total correlation of less than .20, with some items – particularly in the first test task – showing high discrimination at levels around .50. The reading test therefore also demonstrates psychometric properties suitable for a test of this kind.

While both the listening test and the reading test were relatively easy for the population, this does not in itself equate with a deficiency in the instrument. The aim of the *Épreuve Commune* development process was to create a test pitched at the A2 level. Standard setting confirmed that these tests were, indeed, at the A2 level, with the majority of items judged by a panel of judges to correspond with the A2 level descriptors of the CEFR. What this test revealed, then, was that the *Épreuve Commune*-level test-taking population were comfortably answering items at the A2 level, and might in fact have already entered the B1 level. This type of information about learners is only revealed through a test designed through the processes described above, where the criterion is known and verified, where the test is shown to be reliable with high-quality items, and where the test is standardised across sites and administrations. In other words, the *Épreuve Commune* has demonstrated that good testing practice can illuminate vital information about students' levels which is of use for further planning and curriculum design among educational policy makers.

3.1.3 Teacher feedback

Teacher feedback on the 2013 *Épreuve Commune* showed that teachers were, on the whole, very happy with the format and content of the exam (full results are shown in Appendix 1). Although the sample size was relatively small, the teachers showed a clear liking for the materials, with almost all evaluative items receiving a mean score over 4 on a 5-point scale where 5 was “strongly agree”. In the future, it would be worthwhile for the team to administer similar questionnaires to the test-taker population in order to gauge students' views, which are particularly useful in establishing whether instructions are clear, timing is sufficient, and whether or not topics are interesting and motivating.

3.2 The 2014 *Épreuve Commune* administration

3.2.1 Overall test

At the time of writing this report, the initial data on the 2014 *Épreuve Commune* administration was being released. The item level data for the 2014 test were not yet available, thus no detailed analyses are presented here.

Analyses of the overall results of the 2014 *Épreuve Commune* (see Appendix 3) indicate that there was a strong correlation (Cohen, 1988) between the test-takers' results on the test and their year-average result (with the exception of one sub-group of learners with a medium-strength correlation). In addition, the analyses showed that there were no significant differences in performances on the *Épreuve Commune* between test-takers from various sub-groupings.

3.2.2 Teacher feedback

Similar to the 2013 test cycle, teachers' views were collected on the 2014 *Épreuve Commune*. As compared to 2013, a shorter questionnaire was distributed, hoping to receive a greater response rate. Indeed, the response rate doubled. Full results are shown in Appendix 2.

Teachers' views indicated that they found the exam's overall quality, the communication concerning the exam, as well as its ease of administration to be of a good standard; most questionnaire items received a mean score over 5 on a 6-point scale where 6 was "strongly agree". Overall, the teachers also expressed positive views on scoring-related aspects of the exam (with most items receiving a mean score close to 5), although this may be an area to prioritise in further development in future years.

4 Dissemination

Dissemination of materials and communication of the test development activities are key aspects of any successful testing programme. The TDE team have been very proactive in this way, and have connected with teachers and other stakeholders in numerous ways over the past two years in particular. Further details are provided under the headings below.

4.1 Stakeholder exchange

Members from the TDE group (Maurice Broers, Michel Fandel, Halldor Halldorsson, and Romy Schmit) regularly attended the curricular board meetings of the *Commission Nationale des Programmes d'Anglais* in both the ES and EST school systems. Two to three curricular board meetings are held per academic year, and these are attended by a representative from each Luxembourg school who reports back to his/her school. At these meetings, the TDE members informed colleagues about the progress and modalities of the *Épreuve Commune*. They also reiterated the background and purposes of the TDE group in some of those meetings.

4.2 The *Épreuve Commune* website

Early in 2014, the TDE team helped to set up a website with the aim of disseminating materials and information about the test to the broader stakeholder community. The website is hosted by SCRIPT, and can be found at: <http://portal.education.lu/epreuvescommunes/English.aspx>

The site is impressive in its breadth, and contains information on the members of the team, its history, the project aims and ambitions, a testing blueprint, the test specifications, marking tools for teachers, materials from the *Épreuve Commune* in 2012/13 and 2013/14, as well as an overview of the team's work in 2012/13. It also provides more general web-links on language testing as a useful resource for teachers.

4.3 Cascading language assessment literacy

One of the aims of the training part of this programme was to cultivate a sophisticated level of language assessment literacy in a core team, which could then be passed on to others in the Luxembourg context through pre-sessional training and professional development. Examples of these types of activities are:

- a. Michel Fandel and Halldor Halldorsson have been integrating language test development units into their training work at FOPED, the secondary school teacher training programme at the University of Luxembourg. Their own training and experiences on the TDE project have enabled Michel Fandel and Halldor Halldorsson to enhance the sort of training that student teachers receive. They have integrated various theoretical and practical elements from the Lancaster-designed courses, as well as from the development cycle of standardised tests such as the *Épreuve Commune*, into the “evaluation and assessment” modules of the English FOPED teacher training course. In the corresponding seminars and workshops, student teachers have been familiarised with relevant concepts in test design theory (including validity, reliability, and practicality), enabling them to apply more systematic qualitative test item analysis in regard to specific tasks from their own classroom tests, both during teacher training courses and throughout their future teaching practice. They have also gathered experience with a variety of assessment tools (e.g. use of rating scales). Furthermore, important stages of standardised test cycles (e.g. item moderation and revision, rater standardisation) have been explained and simulated within these teacher training workshops. Since September 2012, four successive cohorts of trainee English teachers have taken these evaluation modules (approximately 15-20 student teachers per year).
- b. Romy Schmit has disseminated the conventions of language test design to *chargés d'éducation*² as part of the training organised by the *Institut de Formation Continue*, which is the provider of in-service teacher training in Luxembourg. These courses treated similar foci and topics as the above-mentioned FOPED seminars.
- c. The team has ensured to have a strong presence at the *English Teachers' Days*, an annual event which aims to share and exchange ideas, methods and resources between English teachers in Luxembourg and which is organised by the English Department of FOPED and endorsed by the Ministry of Education.

At the 2012 *English Teachers' Day*, workshops were held on assessing and/or testing extensive and intensive reading of longer texts (Michel Dohn, Halldor Halldorsson, and Marc Trierweiler), and on testing speaking (Michel Fandel, Christiane Klein, and Romy Schmit). The keynote speech was delivered by Professor J. Charles Alderson from Lancaster University on the CEFR.

At the 2013 *English Teachers' Day*, three members of the TDE steering committee (Maurice Broers, Michel Fandel, and Romy Schmit) presented the development cycle and statistical results of the first English *Épreuve Commune*, which had been implemented on a

² *Chargés d'éducation* are supply teachers who receive basic teacher training but have not obtained fully qualified teacher status by passing the state recruitment exams and completing the regular teacher training programme.

national scale earlier that year. An hour-long presentation (see Appendix 4) was delivered to a significant portion of the Luxembourg English language teaching community, reflecting the team's desire and active attempts to work transparently, give their colleagues the opportunity to provide crucial, constructive feedback, and to encourage other teachers to join or help the TDE team in their efforts.

- d. In June 2014 the team conducted a standardisation session using the latest version of the evaluation grids for the writing component of the *Épreuve Commune*. The whole English department of the École Privée Sainte-Anne were involved in a workshop that, apart from enabling the team to standardise the marking grids, aimed at raising assessment literacy in this area.
- e. As members of SCRIPT at the Ministry of Education, TDE team members Maurice Broers and Romy Schmit have also been coordinating the development of other *Épreuve Commune* tests for French, German, and natural sciences. Knowledge gained from the development of the English *Épreuve Commune* was made available to other test design groups and expertise passed on.

4.4 Communicating beyond Luxembourg

While the TDE team have been successful in creating interest in the *Épreuve Commune* and in raising assessment literacy more generally, there is still a need to communicate, and to engage, with the broader testing community beyond Luxembourg. It would be ideal if team members could have more opportunities to connect with networks of language assessors across Europe, and to present their work at conferences, where it would be received with great interest. Drs Brunfaut and Harding will be describing the origins of the project and its current status at the EALTA conference in Copenhagen (Denmark) in May 2015. In addition, a book chapter proposal by Drs Brunfaut and Harding on the team's work has been accepted for an edited volume on teacher involvement in large-scale language testing (Springer). It is believed that a broader network of expert contacts will help to keep the project sustainable.

5 Summary and recommendations

In summary, the TDE team has come a very long way since 2011. They have increased their knowledge and skills in language testing to the point where they can design their own high-quality tests, and also teach others how to construct and evaluate assessments according to the principles of good test design. They have also developed as a cohesive group, working towards deadlines together, solving problems jointly, and managing relationships within the team professionally. The team has so far produced two versions of the *Épreuve Commune*, and the empirical evidence demonstrates that the team is capable of developing a high-quality test product on par with many commercial products available across Europe. The three years of the TDE project thus far has seen the emergence of a team of language testers who now possess skills and knowledge in language assessment which are of an international standard, who can develop in-house, high-quality assessments suitable for the Luxembourg context, and who therefore should be viewed as a highly prized group of professionals within the landscape of language education in Luxembourg.

In taking the team further, and in ensuring that the project remains sustainable, we would make several recommendations with regard to the team's immediate needs:

1. The TDE team requires more support in terms of time out of the classroom. Compared with similar exam reform projects in Austria and Hungary, for example, the TDE team do not have enough available time to capitalise on their potential. It would be useful if at least some members of the team could receive more than one day of relief per week. This would provide more time to act on the various tasks that lie ahead in continuing to develop the *Épreuve Commune*, more time to sufficiently prepare the specifications for a new school-leaving exam, and more time for members of the team to engage with teacher networks within Luxembourg, and with the professional community of language testers across Europe. Currently, many members of the team work through weekends and holidays to keep the project viable, without additional compensation. This is not a sustainable practice, and if the quality and reliability of assessment in the Luxembourg school system is to continue to improve, the TDE team need to be supported with greater time resources.
2. As mentioned above, there is also the question of financial resources. Team members can not currently attend important conferences to share their research and ideas because there is no means of financing this. There is also no money available for the sort of research work which would be necessary for the ongoing validation of a high-quality testing programme. Investment in the team at this stage would be crucial in ensuring the ongoing sustainability of the project.
3. The team was originally put together with a view to designing a new secondary school leaving exam. In a sense, the *Épreuve Commune* was the training ground for this main project. The members of the team have now shown themselves to be capable of producing high-quality tests which adhere to international standards of practice. It would be fruitful if the team could now be supported, and indeed actively encouraged, to take the project to the next stage and begin to outline specifications for a school-leaving exam. While we are aware that the team is already making progress on this, unless points 1 and 2 above are addressed it is likely that the reform will not be sustainable.

6 References

- Alderson, J.C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Brunfaut, T., & Harding, L. (2014). *Linking the GEPT listening test to the Common European Framework of Reference*. Research Report RG-05. Taiwan: Language Training and Testing Centre. <https://www.lttc.ntu.edu.tw/lttc-gept-grants/RReport/RG05.pdf>
- Cohen, J. (1988). *Statistical power analysis for behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Council of Europe (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Council of Europe (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. A manual*. Strasbourg: Language Policy Division. http://www.coe.int/t/dg4/linguistic/source/manualrevision-proofread-final_en.pdf
- Green, A. (2014). *Exploring language assessment and testing*. Oxon, UK: Routledge.
- Green, R. (2013). *Statistical analyses for language testers*. Basingstoke, UK: Palgrave Macmillan.

7 APPENDICES

7.1 Appendix 1: 2013 *Épreuve Commune* teacher questionnaire data

Note: The questions were judged on a 5-point Likert scale (1=strongly disagree, 5=strongly agree)

	<i>M</i>	<i>SD</i>	Min	Max	<i>n</i>
Listening Paper					
The materials (recordings, input texts, instructions, tasks ...) are free of errors.	5.00	.00	5	5	17
The recordings are clear and intelligible.	4.12	1.05	1	5	17
The length of the paper is appropriate for the time available.	4.47	.72	3	5	17
The length of the paper, the number of tasks and questions are appropriate to gain an impression of the level reached by the pupils.	3.76	1.03	2	5	17
The paper reflects the contents of the English course of the pupils.	4.24	.66	3	5	17
The paper focuses on important content rather than incidental information (trivia, exceptions ...).	4.47	.62	3	5	17
The topics of the input texts are sufficiently varied to cover the content areas specified in the syllabus and in the specifications.	4.65	.49	4	5	17
The topics of the input texts are appropriate in terms of the age, the interests and the general knowledge of the Pupils.	4.24	.75	3	5	17
The pupils are familiar with the types of tasks in the paper.	4.29	1.31	0	5	17
The types of tasks are sufficiently varied not to give an unfair advantage to certain pupils.	4.71	.47	4	5	17
The instructions are clear and complete.	4.82	.53	3	5	17
The questions are worded and presented clearly.	4.94	.25	4	5	16
The degree of difficulty of the paper, the tasks and the questions is appropriate.	4.00	1.17	2	5	17
Reading Paper					
The materials (input texts, instructions, tasks ...) are free of errors.	5.00	.00	5	5	17
The length of the paper is appropriate for the time available.	4.56	.78	2	5	18
The length of the paper, the number of tasks and questions are appropriate to gain an impression of the level reached by the pupils.	4.39	.70	3	5	18
The paper reflects the contents of the English course of the pupils.	4.50	.62	3	5	18
The paper focuses on important content rather than incidental information (trivia, exceptions ...).	4.56	.78	2	5	18
The topics of the input texts are sufficiently varied to cover the content areas specified in the syllabus and in the specifications.	4.44	.70	3	5	18
The topics of the input texts are appropriate in terms of the age, the interests and the general knowledge of the pupils.	4.50	.86	2	5	18
The pupils are familiar with the types of tasks in the paper.	4.44	.92	2	5	18
The types of tasks are sufficiently varied not to give an unfair advantage to certain pupils.	4.78	.43	4	5	18

	<i>M</i>	<i>SD</i>	Min	Max	<i>n</i>
The instructions are clear and complete.	4.94	.24	4	5	18
The questions are worded and presented clearly.	4.94	.24	4	5	17
The degree of difficulty of the paper, the tasks and the questions is appropriate.	4.33	.84	3	5	18
The materials (prompts, instructions, tasks ...) are free of errors.	4.94	.24	4	5	18
Writing Paper					
The length of the paper is appropriate for the time available.	4.28	1.27	1	5	18
The length of the paper, the number of tasks and questions are appropriate to gain an impression of the level reached by the pupils.	3.67	1.03	2	5	18
The paper reflects the contents of the English course of the pupils.	4.22	.88	3	5	18
The paper focuses on important content rather than incidental information (trivia, exceptions ...).	4.17	.62	3	5	18
The topics to write on are appropriate in terms of the age, the interests and the general knowledge of the pupils.	4.22	.73	3	5	18
The pupils are familiar with the types of tasks in the paper.	4.78	.55	3	5	18
The types of tasks are sufficiently varied not to give an unfair advantage to certain pupils.	4.28	.89	2	5	18
The prompts, instructions are complete.	4.67	.84	2	5	18
The prompts, instructions are worded and presented clearly.	4.83	.51	3	5	18
The degree of difficulty of the paper and the tasks is appropriate.	3.89	1.13	2	5	18
Overall Test					
The test booklet is clear, legible and attractive.	4.83	.38	4	5	18
The overall variety and range of the test is sufficient to gain an impression of the level reached by the pupils.	3.94	.94	2	5	18
The overall length of the test is appropriate to gain an impression of the level reached by the pupils.	4.56	.51	4	5	18
The overall length of the test is appropriate in terms of the pupils' capacity to work in a concentrated manner and perform well.	4.11	1.02	1	5	18
The test reflects the contents of the English course of the pupils.	4.33	.91	2	5	18
Administration					
The teachers, invigilators have been given enough prior information to administer the test efficiently.	4.00	1.03	2	5	18
The pupils have been given enough prior information to sit the test successfully.	4.06	1.00	2	5	18
The test is easy to administer.	4.67	.59	3	5	18
The test can be administered under conditions that give all pupils an optimal chance – confusion or disturbance	4.28	.75	3	5	18
Cheating can be prevented easily.	3.71	1.05	2	5	17
Most pupils finish in the time allotted.	4.50	.71	3	5	18
Marking Tools					

	<i>M</i>	<i>SD</i>	Min	Max	<i>n</i>
The answer keys for listening and reading are complete, free of errors and clear.	4.72	.57	3	5	18
The marking procedures for listening and reading are clear and fair.	4.61	.70	3	5	18

7.2 Appendix 2: 2014 *Épreuve Commune* teacher questionnaire data

Note: The questions were judged on a 6-point Likert scale (1=strongly disagree, 6=strongly agree)

	<i>M</i>	<i>SD</i>	Min	Max	<i>n</i>
Overall Test					
The materials (recordings, input texts, instructions, tasks ...) are free of errors.	5.8	.50	5	6	39
The recordings for the listening tasks are clear and intelligible.	4.6	1.40	2	6	38
The prompts / instructions are clear and complete.	5.7	.50	4	6	39
The tests booklets are legible and attractive.	5.7	.70	3	6	39
The pupils are familiar with the types of tasks in the paper.	5.1	1.10	2	6	39
The length of the test is appropriate for the time available.	5.2	1.20	2	6	40
The length of the test, the number of tasks and questions are appropriate to gain an impression of the level reached by the pupils.	4.7	1.20	1	6	40
Pupils who have been taught the syllabus should be ready to perform well in the test.	5.1	.90	3	6	40
The degree of difficulty of the test is appropriate.	4.4	1.50	1	6	40
Administration					
There is enough information for teachers /invigilators to administer the test efficiently.	5.4	1	3	6	40
There is enough information for pupils to sit the test successfully.	5.3	.90	3	6	40
The test is easy to administer.	5.4	.80	2	6	40
The test can be administered under conditions that give all pupils an optimal chance – confusion or disturbance that could interfere with effective performance are avoided.	5.1	.80	3	6	40
Marking					
The listening and reading papers are easy to mark.	5.6	.90	5	6	40
The marking procedures for writing are clear and fair.	4.8	1.00	2	6	40
Associating bands to the pupils' written performances is easy.	4.4	1.1	2	6	40
The marked sample tasks provided in this booklet are useful.	4.9	.90	3	6	39
The marking grids are useful for giving feedback to the pupils.	4.7	1.00	4	6	39
Results					
The marks are close to the marks the pupils achieve in in-class tests.	4.1	1.40	1	6	39
The marks reflect the teacher's expectations.	4.6	1.00	1	6	39

7.3 Appendix 3: Initial analyses of the 2014 Épreuve Commune

Epreuves Communes 2014 English Results

Samples (n):

Version PO

- 63 pupils from 2 Schools in 3 classes

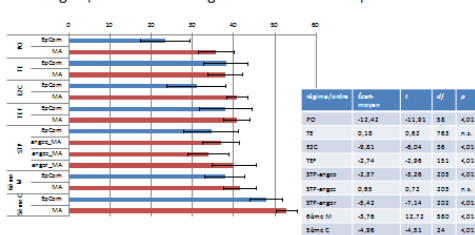
Version TE

- E2C: 37
- STP: 205
- TE: 765
- TEF: 152

6ème M: 581

5ème C: 25

Average EpCom scores and grades: Individual comparisons

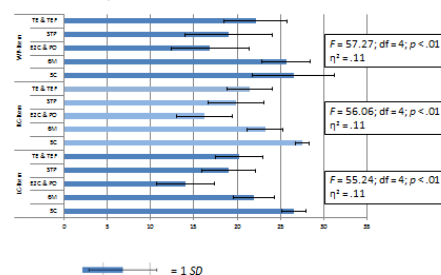


Correlations between test scores and average grades

EPCOM \longleftrightarrow r Moyenne annuelle

régime/ordre	Pearson r
PO	.74
TE	.68
E2C	.96
TEF	.79
STP-angco	.63
STP-angec	.42
STP-angor	.61
6ème M	.69
5ème C	.72

Raw-score comparisons



Raw-score comparisons: Homogeneous Groups

LC-Rem						RC-Rem					
Tukey HSD ^{a,b}						Tukey HSD ^{a,b}					
track	N	1	2	3	4	track	N	1	2	3	4
E2C & PO	96	14.08				E2C & PO	96	18.23			
STP	205	18.00				STP	205	19.85			
TE & TEF	917	20.28	20.28			TE & TEF	917	21.47	21.47		
6M	581		21.92			6M	581		23.22		
SC	25			26.52		SC	25			27.51	
Sig.		1.000	.533	.263	1.000	Sig.		1.000	.228	.161	1.000

Means for groups in homogeneous subsets are displayed.
 a. Uses Harmonic Mean Sample Size = 88.048.
 b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

WP-Rem					
Tukey HSD ^{a,b}					
track	N	1	2	3	
E2C & PO	96	16.88			
STP	205	18.99			
TE & TEF	917		22.11		
6M	581			25.84	
SC	25				28.40
Sig.		.316	1.000		.884

Means for groups in homogeneous subsets are displayed.
 a. Uses Harmonic Mean Sample Size = 88.048.
 b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Raw-score comparisons: Wrap up

- No significant differences between STP and TE for LC and RC
- No significant differences between TE and 6M for LC and RC
- No significant differences between E2C & PO and STP for WP
- No significant differences between 6M and 5C for WP

7.4 Appendix 4: 2013 English Teachers' Day presentation

TDE-PRESENTATION

English Teachers' Day
17 October 2013

TDE @ ETD
INTRODUCTION

WHO WE ARE

- TDE = Test Design and Evaluation
- In existence since 2011/12
- Group of 9 teachers from ES, EST and INL
- Idea for group originated within CNP-ES
→ Desire to improve school-leaving exams
- Collaboration with Lancaster University, SCRIPT and uni.lu/FOPED

WHAT WE AIM FOR

- Building up local expertise about test design and assessment
 - 3 taught modules by Lancaster University experts over the course of 3 years
 - External consultancy supporting the professional development of local tests
- Sharing and multiplying that knowledge within the Luxembourg teaching community
- Proposing ways of designing school-leaving exams
 - In line with current international standards
 - Focusing on specificities and requirements of Luxembourg context

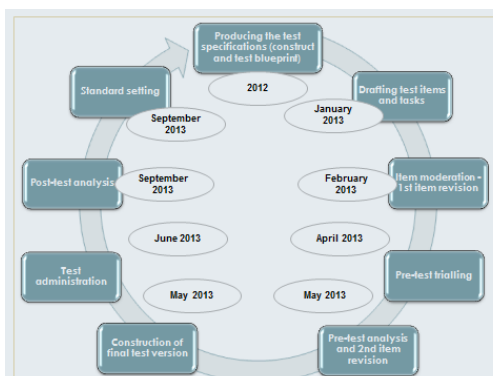
WHAT WE DO

- 3-year course in language test design and evaluation
 - Final module of taught course this year
- Turning theory into practice
 - English *Epreuve Commune* (6eM, 5eC, 9eTE) = test developed according to recommended test cycle
- Developing and sharing test and assessment tools
- Cooperation with the Luxembourg ELT community, CNP-ES/EST, FOPED...

WHAT WE DON'T DO

- Curriculum and syllabus design
 - Our focus = how to test, not what to teach
- "Just" the *Epreuve Commune*
 - Tests at other levels envisaged
 - Our long-term goal = improvement of school-leaving exams

THE TEST DEVELOPMENT CYCLE



THE TEST BLUEPRINT

2013/2014	Time	Items / tasks per skill	Marks
Listening	30'	30	20
Reading	30'	30	20
Writing	40'	2 tasks	20
Total	100'		60

2012/2013	Time	Number of tasks	Items per skill	Task	Items per task	Test methods
Listening	30'	4	30	L1	6+2	5 items, 4-option MCQs with pictures and 2 items in a 6-option checklist
				L2	4	3-option MCQs or MM (4 answers + 2 distractors)
				L3	8	TFDS, sentence or table completion or 3-option MCQs (prompt = monologue)
				L4	10	TFDS, sentence or table completion or form filling (prompt = dialogue)
Reading	30'	4	30	R1	7	MM (7 answers + 2 distractors) or information transfer
				R2	7	MM (signs and statements + 3 distractors), 3-option MCQs or 4-option MCQs
				R3	8	3-option MCQs, TFDS, short answers, banked gap-fill
				R4	8	3-option MCQs, TFDS, short answers, banked gap-fill
Writing	40'	2	n/a	W1	n/a	Guided writing (interactive writing), 70-80 words
				W2	n/a	Free writing (writing a narrative), 120-140 words
Total	100'	10				

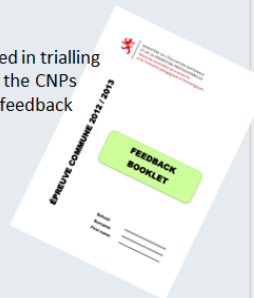
ANALYSIS OF THE TEST VERSION 2012/2013

A FEW STATISTICS

Qualitative Analysis

Feedback

- from teachers involved in trialling
- from exchanges with the CNPs
- through a dedicated feedback questionnaire



MARKING THE WRITTEN PRODUCTION – WRITING PAPER

Marking grids

- The descriptors of the marking grids describe a logical progression in terms of language learning.

strongly agree

5	4	3	2	1	0
---	---	---	---	---	---

 strongly disagree

- The descriptors encompass the various levels of performance displayed by the majority of pupils.

strongly agree

5	4	3	2	1	0
---	---	---	---	---	---

 strongly disagree

Statistical Analyses

Sample

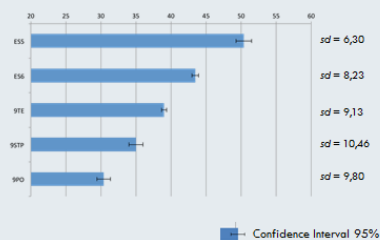
In total, 1800 students participated in the English EpCom in 2013. Of those, we had data for the following sample:

$n = 1.216$ students in 63 classes of 19 schools:
EPF, EPMC, LCD, LCE, LEM, LGL, UBM, LMRL, LN, LNB, LTAM, LTE, LTJB, LTL, LTMA, LTML, NOSL, SLL, UELL

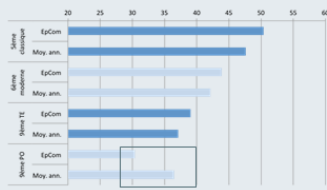
5^{ème} classique: $n = 33$
6^{ème} moderne: $n = 322$
9^{ème} TE: $n = 646$
9^{ème} STP: $n = 110$
9^{ème} PO: $n = 104$

MARKS OBTAINED IN THE 'EPREUVE COMMUNE'

Differences between the various school types and streams

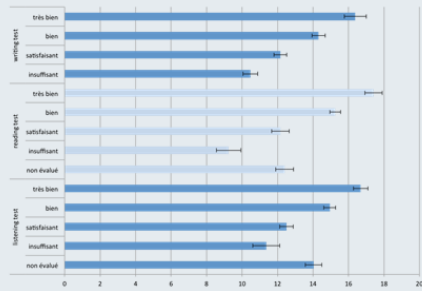


Comparison of students' 'Epreuve Commune' marks and their 'annual average' in English



ANOVA with two factors

- Repeated measurement factor (EpCom – MA angl)
 $F = 0.05$, $p = .85$, $\eta^2 = .00$
- Factor 'School types and stream'
 $F = 65.45$, $p < .01$, $\eta^2 = .16$
- Interaction factor (repeated measurement (EpCom – MA angl) * 'School types and streams')
 $F = 35.51$, $p < .01$, $\eta^2 = .09$



Analyses for PROCI classes

Spearman Correlation coefficients:
 $-1 \leq r \leq 1$
 $r = (-)$ 1 linear relationship between variables
 $r = 0$ no relationship between variables

EpCom	Evaluation in PROCI classes			
	Comprehension	Ecrit	Oral	
Listening Test	0,67	0,30	0,27	
Reading Test	0,70	0,48	0,29	
Writing Test	0,59	0,61	0,52	
Total Mark	0,79	0,57	0,43	

$n = 110$

Statistical Analyses Wrap up

- strong differences between school types and streams
- strong correlations between the students' EpCom results and the marks they obtain in their every day tests
- PO students systematically obtained lower EpCom results compared to their usual grades
- strong correlations between teachers' evaluations and EpCom results among PROCI-students



Note:

All this information and last year's test will be available online before Christmas.

WHAT YOU CAN DO TO HELP US

- Feedback about our products
 - Invaluable and indispensable
 - E.g. *Epreuve Commune* feedback booklets / CNP feedback
- Active participation in TDE test projects / at various test stages
 - Item writing
 - Item moderation
 - Recording of listening tasks
 - Standard setting
 - Trialling / pre-testing

What is the added value?

- Additional perspective on the students' work
- Outside view based on objective assessment
- Clear-cut information on the students abilities in the different skills
- Illustration of the A2 level
- Confirmation of the quality work done by the English teachers

THANK YOU FOR YOUR ATTENTION!

Enjoy the rest of ETD 2013!