Study on comparability of language testing in Europe

Final report
September 2015

Education and Training

Prepared by:
CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

EUROPEAN COMMISSION

Directorate-General for Education and Culture

Directorate A — Modernisation of Education I: Europe 2020, country analysis, Erasmus+ coordination

Unit A.4 — Studies, impact assessments, analysis and statistics

E-mail: eac-unite-a4*@ec.europa.eu*

European Commission
B-1049 Brussels

EUROPEAN COMMISSION

Study on comparability of language testing in Europe

Final report

September 2015

Directorate-General for Education and Culture

2015 Erasmus+ **EN**

Europe Direct is a service to help you find answers to your questions about the European Union.

Freephone number (*)**:**

00 800 6 7 8 9 10 11

(*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

More information on the European Union is available on the Internet (http://europa.eu).

Luxembourg: Publications Office of the European Union, 2015

ISBN 978-92-79-50995-7

doi: 10.2766/082033

Printed in Belgium

Table of Contents

Abstract

Following the "Conclusions on Multilingualism and the Development of Language Competences", adopted in May 2014, the Council of the European Union invited the European Commission to explore the feasibility of assessing language competences across all the Member States by making use of existing national language tests. This study looked at 133 national language examinations (33 jurisdictions, 28 EU Member States) at ISCED 2 and ISCED 3 levels. The languages included were EU official languages other than the main language of instruction which are studied by more than 10% of the students in secondary education in each jurisdiction. This study adopted a mixed methods approach which included the analysis of qualitative data – collected through the expert content analysis of the examinations – and of quantitative data – collected through a comparative judgement exercise. The results from this study show that the meaningful comparability of national results of language examinations across EU Member States in the future will depend on 1) the results being expressed in a uniform format; 2) implementing measures at both national and European level that would increase the quality of current language examinations, and in turn ensure that results are similarly valid and reliable across all jurisdictions.

Après les "Conclusions sur le Multilinguisme et le Développement des Compétences linguistiques" adoptées en mai 2014, le Conseil de l'Union Européenne a invité la Commission Européenne à explorer la faisabilité de l'évaluation des compétences linguistiques en s'appuyant sur les tests linguistiques nationaux existants dans les États Membres de l'UE. Cette étude comprend l'analyse de 133 examens nationaux de langues (33 territoires, 28 États Membres) aux niveaux CITE 2 et CITE 3. Les langues incluses dans l'étude sont les langues officielles de l'UE autres que les langues d'instruction et étudiées par plus de 10% des élèves de l'enseignement secondaire dans chaque territoire. Cette étude a adopté une méthode de recherche mixte comprenant l'analyse des données qualitatives – rassemblées par une analyse experte des examens – aussi que quantitatives – rassemblées par un exercice de jugement comparatif. Les résultats de cette étude montrent que la comparaison des résultats nationaux des examens de langues des États Membres dépendra 1) du format uniforme de ces résultats; et 2) de la mise en pratique des mesures nationales et européennes visées à une augmentation de la qualité des examens linguistiques actuels et, par la suite, à assurer que les résultats soient similairement valables et fiables dans tous les territoires.

1 Executive Summary

1.1 Purpose and scope of the study

Following the "Conclusions on Multilingualism and the Development of Language Competences", adopted by the Council of the European Union in May 2014, a new approach was suggested for measuring language competences at the European level. Rather than develop a language benchmark across all Member States, it was concluded that measures should be implemented for promoting multilingualism and enhancing the quality and efficiency of language learning and teaching, and to develop measures for assessing language proficiency preferably within each country's educational system.

To develop an evidence base and understanding of language competences in Europe, the Council invited the European Commission to explore the feasibility of assessing language competences across all the Member States by making use of existing national language tests. The aim of this study is to critically assess the comparability of existing national tests of pupils' language competences in Europe at both ISCED 2 and ISCED 3 levels. The study draws upon data on existing national tests of language competences in the 28 EU Member States collated by the Eurydice Network.

1.2 Languages and examinations included in the study

The languages included in this study are languages that are not the main language of instruction. Only EU official languages that are used in at least one other EU Member State were included in this study. For each jurisdiction, only those languages studied by more than 10% of secondary education students (according to Eurostat; 2013, 2014) were considered.

On the basis of the data collected by Eurydice, 133 national language examinations (33 jurisdictions, 28 EU Member States) were identified as relevant for this comparability study. Out of these 133 language examinations, 77 were at ISCED 2 level and 56 were at ISCED 3 level. Appendix 1 offers a detailed list of the national exams included in this study, as well as the reasons why certain exams had to be excluded.

1.3 Participation of Member States

In order to ensure that the results of this study are as accurate and transparent as possible, the European Commission facilitated the collaboration of the members of the Indicator Expert Group on Multilingualism (IEG). These members are all experts in language education and/or language assessment working for the Ministries of Education or National Statistical Offices in their respective jurisdictions.

After an initial meeting with the European Commission and the above-mentioned group of experts, the Project Team established direct contact with each of the members of the group to discuss in more detail the national language tests existing in each jurisdiction. The members' contribution was key to confirm the languages and tests chosen for each jurisdiction, and to provide any additional information regarding the exams (test papers, samples of students' performance, supporting documentation regarding the tests e.g., procedures for the creation and administration of exams, training materials for item writers and raters, national results, etc.).

1.4 Structure of the study

The five main tasks considered by this report are:

- Task 1: Assessment of comparability of the existing national language tests administered to secondary school students.
- Task 2: Proposals for ex-post adjustment that can increase the comparability of existing results.
- Task 3: Proposals for development work that can increase comparability of existing language tests.
- Task 4: Proposals for Member States not having a system for language testing and interested in developing one.
- Task 5: Comparative overview of existing country data on language testing

The Common European Framework of Reference (CEFR) was used as the comparative framework in this study. The CEFR is very widely used throughout Europe and serves as a familiar point of reference, a relevant model of language learning, and a measurement construct.

1.5 Findings

Task 1 above was conducted using a mixed methods approach which included the analysis of both quantitative and qualitative data, and which is described in detail in section 5. The qualitative data was collected through the expert content analysis of existing language examinations by a group of highly-competent specialists in language assessment from across Europe. These experts used an online content analysis tool and were specifically trained on the use of this tool to ensure the consistency and reliability of their work. The quantitative data was collected through a comparative judgement exercise which was conducted by 49 experts in language education and assessment on an online platform designed for this purpose (www.nomoremarking.com).

The qualitative ***content analysis*** of test features looked at 133 language examinations (33 jurisdictions, 28 EU Member States). Considerable diversity was found across these language examinations, which decreases the potential for a straight-forward comparison of test results. Four main areas were investigated: constructs (what is measured by the test), the interpretations given to test results, test taking populations, and measurement characteristics (contextual features which may affect comparability). Over a wide range of points, evidence was found which suggests a lack of comparability.

In regards to **constructs**, language examinations from across different jurisdictions show considerable diversity, despite components usually being referred to in the same terms (e.g. 'Reading'). As a consequence of this, it is probably mistaken to compare results of different tests and conclude that they are interchangeable when they are actually testing different constructs. In other words, different tests aim to test different abilities even if they use common terms to refer to the elements tested.

Considering **interpretations of results**, the main finding concerned those tests which did not claim alignment to the CEFR. It was not possible to establish how test results were to be interpreted in many cases. Some interpretations were norm-referenced (to be interpreted by comparing the placement of a candidate to that of

his/her peers). Such an approach is not directly conducive to comparing results between different tests, as the populations in each case would be different.

The **populations** of ISCED 2 and ISCED 3 tests were found to be reasonably homogeneous in respect of age, the only population characteristic examined.

In terms of **measurement characteristics**, as with construct, many of the findings suggested limits on comparability. For example, a significant proportion of tests were not able to demonstrate equivalence across administrations. In this case, comparability of these tests with other tests is impossible because the results of one session cannot even be compared to those of another session for the same test. Although comparability of results between sessions is desirable for a great many reasons, and should be addressed, tests were also diverse for quite legitimate reasons. For example, the item type used has an effect on test result which relates to the nature of the construct, and some types can have a number of unique effects, such as increasing or decreasing the discrimination between candidates.

A quantitative approach to comparing existing results using ***comparative judgement*** was also presented, and illustrated with a limited sample of Reading and Writing tasks from the language examinations included in this study. This method shows how national results of the different jurisdictions can be aligned to the CEFR on the basis of the difficulty of the tasks in their different national language exams. This study was able to demonstrate differences in the relative difficulty of tasks across language examinations, but due to the limited scope of the study it was not possible to provide a full comparison of the results of individual tests as data concerning score distributions was in most cases unavailable. Given the current lack of direct comparability between national test results, the method presented suggests a new way in which results of national test could be compared in the future, especially if the comparative judgement technique was applied to the samples of students' performance in Writing and Speaking tasks.

1.6 Proposals for development

In view of the findings from Task 1, a number of proposals were put forward in order to address Task 2, Task 3 and Task 4. The following proposals are explained in detail in sections 6, 7 and 8.

1.6.1 Proposals for ex-post adjustment to increase the comparability of existing national results

This study suggests the use of comparative judgement as the most suitable methodology for ex-post adjustment of existing results. This method aims to build a common scale of language proficiency to which national language exams and results of all jurisdictions could be mapped. However, in order to fully implement this methodology, a number of conditions need first to be met:

- A common approach to reporting national results
- Jurisdictions' commitment to provide relevant evidence
- An annual schedule set and monitored by a responsible body

1.6.2 Proposals for development work to increase the comparability of existing language tests

The extent to which test results are comparable is affected by test quality and by diversity due to legitimate differences in testing contexts and purposes. Test quality affects comparability because weaker, less reliable measurement leads to unreliable results. The findings of this report show that there are a number of quality issues affecting tests which should be addressed by national assessment boards. It should be recognised, however, that some improvements may be constrained in some jurisdictions by a number of factors, such as costs or educational context. Lack of comparability due to legitimate differences between tests is harder to mitigate, and cross-jurisdiction comparability would need to be incorporated as an aim in each case. The main recommendations for review and possible implementation are therefore:

Construct

- expand the range of the types of reading and listening tested at B2 and above;
- design tasks which elicit the appropriate cognitive processes for each CEFR ability level.

Interpretations

- develop criterion-based interpretations of test results which may be mapped to the CEFR if alignment to the CEFR is not to be sought.

Population

- collect information regarding the characteristics of those taking the test.

Measurement Characteristics

- ensure that recruitment of all staff (test developers, item writers, editors, markers, raters, analysts, etc.) is based on the full set of competences required for the job;
- ensure that deficiencies in staff competences is addressed by training;
- ensure that rater judgement is standardised so that consistent judgements are made;
- ensure rating procedures involve monitoring and remedial action in cases where the monitoring reveals issues;
- develop procedures to correct for differences (especially in difficulty) between forms of the same test;
- pursue a thorough programme which aims to align the test to the CEFR;
- routinely collect score and response data and analyse it to initiate improvement in procedures of development and administration;
- improve item writing and editing processes to remove item flaws;
- review legitimate features of the test and determine whether they can be made more comparable with those of tests from other jurisdictions;
- consider the use of a single test for comparison of candidate ability across jurisdictions.

1.6.3 Proposals for the development of future national language examinations

There exists extensive literature with theoretical and practical recommendations for the effective design and implementation of language examinations, and these have

been referred to in section 8. Beyond these general guidelines, a number of concrete recommendations were suggested due to their potential impact on the comparability of future results of national language examinations.

- Design the CEFR into the test: the task of designing tests based on the CEFR will be easier if the CEFR is used as the starting point.
- Develop procedures to continually improve the test: test provision must be seen as a cycle where information is continually gathered in an attempt to detect issues and resolve them for future tests.
- Develop a process to maintain standards: setting where the boundaries are between CEFR levels should be done once and then the standards should be maintained thereafter, preferably through item banking.

1.7 Comparative overview of existing country data on language testing

Task 5 required providing an overview of the data that is currently available from all jurisdictions regarding language test results. The focus of this task was only on results for the first foreign language in each jurisdiction, and the data should preferably come from publicly available sources.

Out of the initial 133 language examinations included in this study, we attempted to collect data for 62 tests of first foreign languages from 33 jurisdictions, but could only find relevant data for 45 of these tests from 26 jurisdictions. The reasons why results may not be available are described in section 9.

Data available differed greatly from jurisdiction to jurisdiction, and so did the format in which this information was provided. Section 9.2 presents a summary of the observations made regarding the current format in which national results of language tests are reported.

In order to produce in the future a European summary table of adjusted national results which could be used to regularly monitor students' proficiency in one or several foreign languages, a number of elements need to be carefully considered beforehand to ensure that this table will be compiled and interpreted in the most meaningful and representative way. These elements are explained in more detail in section 9.3, and include the selection of the data that is to be reported, the meaning of "passing" grades, and the test population.

1.8 Conclusion

The extent to which results of national language examinations can be compared depends on a number of factors. First of all, comparisons of national results are only feasible when the data being compared have sufficient elements in common. From the review of this data, there seems to currently exist too much variability on the information made available by the different jurisdictions and the format in which this information is provided. However, and most importantly, this study has shown that language examinations across jurisdictions present a wide variety of features in terms of the constructs tested, the populations of test takers, the interpretations of the results and the measurement characteristics of these examinations. These features importantly determine test quality, and in turn impact on the validity and reliability of the results obtained. **The meaningful comparability of national results of language examinations across EU Member States will therefore depend not**

only on these results being expressed in a uniform format, but also on implementing measures at both national and European level that would increase the quality of current language examinations, and in turn ensure results are similarly valid and reliable across all jurisdictions.

2 Introduction

Europe is a geographical territory marked by its rich diversity at all possible levels, from landscapes to cultures and languages. In an effort to maintain this diversity while encouraging mutual understanding, as the EU motto states "united in diversity", the European Council concluded in Barcelona in 2002 that all European citizens should aim to learn at least two foreign languages. In order to monitor this progress, the Council asked the European Commission to prepare a proposal for a European Indicator of Language Competences (Council Conclusions May 2006 and 2009). As part of these efforts, the Commission undertook the first European Survey on Language Competences (ESLC), completed by the consortium SurveyLang and lead by Cambridge English Language Assessment. The results of this study were presented in 2012 and showed how the European diversity mentioned above was also reflected in the way language learning, teaching and assessment is understood and conducted in the different EU Members States.

Partly in the light of these results, in May 2014 the Council of the European Union rejected the Commission's proposal to create a European benchmark on languages (stated in *Rethinking Education Communication*, 2012) and instead invited the European Commission to explore the extent to which national systems of data collection regarding language proficiency could be compared across jurisdictions. The current study is one of the two responses of the Commission to the Council's invitation, the other response being the inventory on existing national language tests that the Eurydice Network has compiled and that has served as starting point for this study.

2.1 The context of this study

Following the *Conclusions on Multilingualism and the Development of Language Competences*, adopted by the Council of the European Union in May 2014, a new approach was suggested for measuring language competences at the European level. Rather than develop a language benchmark across all Member States, it was concluded that measures should be implemented for promoting multilingualism and enhancing the quality and efficiency of language learning and teaching, and to develop measures for assessing language proficiency preferably within each country's educational system.

To develop an evidence base and understanding of language competences in Europe, the Council invited the European Commission to explore the feasibility of assessing language competences across all the Member States by making use of existing national language tests. These tests are generally organised at the national or regional level, implemented in secondary schools, and funded by national or regional budgets for education and training. To support this initiative, measures will also have to be put in place to encourage member states to develop appropriate methodologies for assessing language proficiency and to adopt a harmonised methodology to enable results to be compared between Member States and at the European level.

2.2 The scope of this study

The aim of this study is to critically assess the comparability of existing national tests of pupils' language competences in Europe at both ISCED 2 and ISCED 3 levels. The study draws upon data on existing national tests of language competences in the EU

Member States collated by the Eurydice Network. In analysing the data, the study aims to determine the extent to which existing national language examinations share enough features to render the results comparable across jurisdictions. Furthermore, an approach to making such comparisons is to be developed.

2.2.1 Definitions

The terms and definitions used in this report have been taken directly from the Tender Specifications, which clearly determined the scope and nature of this study.

By **national language tests** is meant exams 'generally organised at the national or regional level, implemented in secondary schools and funded by national or regional budgets for education and training' (p.14). In most cases, these exams are organised by central/top level public authorities.

However, some EU Member States such as the United Kingdom and Belgium have independent educational authorities for each part/region with different language testing systems. For the purpose of this study, we have considered these parts/regions as separate entities and therefore looked at existing language exams in the 33 educational **jurisdictions** identified across the 28 EU Member States.

The **languages** included in this study are 'languages that are not the main language of instruction; or; in other words, [competences] in one or more foreign languages' (p.14). Considering the current focus on increasing mobility and access to jobs in other EU countries, only EU official languages that are used in at least one other EU Member State were included in this study. For each jurisdiction, only those languages studied by more than 10% of secondary education students (according to Eurostat; 2013, 2014) were considered. In the cases of jurisdictions with more than one official language, DG EAC were consulted and helped determine, in agreement with the relevant Indicator Expert Group on Multilingualism (IEG) members, which language tests should be looked at from each jurisdiction. Table 1 shows the languages selected for each jurisdiction.

Table 1 Foreign languages most taught in each jurisdiction and included in this study

Jurisdiction	First foreign language	Second foreign language	Third foreign language	Fourth foreign language
Austria	English	French		
Belgium FR	English	Dutch	German	
Belgium GE	French			
Belgium NL	French			
Bulgaria	English	German		
Croatia	English	German	Italian	
Cyprus	English	French	Italian	
Czech Rep	English	German		
Denmark	English	German		
Estonia	English	German		
Finland	English	Swedish	German	
France	English	Spanish		
Germany	English	French		
Greece	English	French	German	

Jurisdiction	First foreign language	Second foreign language	Third foreign language	Fourth foreign language
Hungary	English	German		
Ireland	French	German	Spanish	
Italy	English	French	Spanish	
Latvia	English	German		
Lithuania	English	German		
Luxembourg	German	French	English	
Malta	English	Italian	French	
Netherlands	English	German	French	
Poland	English	German		
Portugal	English	French	Spanish	
Romania	English	French		
Slovakia	English	German		
Slovenia	English	German		
Spain	English	French		
Sweden	English	Spanish	German	French
UK England	French	German	Spanish	
UK Northern Ireland	French	German		
UK Scotland	French	German		
UK Wales	French	German		

2.2.2 Examinations included

On the basis of the data collected by Eurydice, 133 language examinations (33 jurisdictions, 28 EU Member States) were identified as relevant for this comparability study. Out of these 133 language examinations, 77 were at ISCED 2 level and 56 were at ISCED 3 level. Table 2 Total number of exams per ISCED level and number of exams testing each of the languages included in the study. shows the number of exams testing each of the languages in this study. Appendix 1 offers a detailed list of the national exams included in this study, as well as the reasons why certain exams had to be excluded.

Table 2 Total number of exams per ISCED level and number of exams testing each of the languages included in the study.

	English	French	German	Spanish	Italian	Swedish	Dutch	Total
ISCED 2	29	15	24	6	2		1	77
ISCED 3	23	13	15	3	1	1		56

2.3 Practical considerations

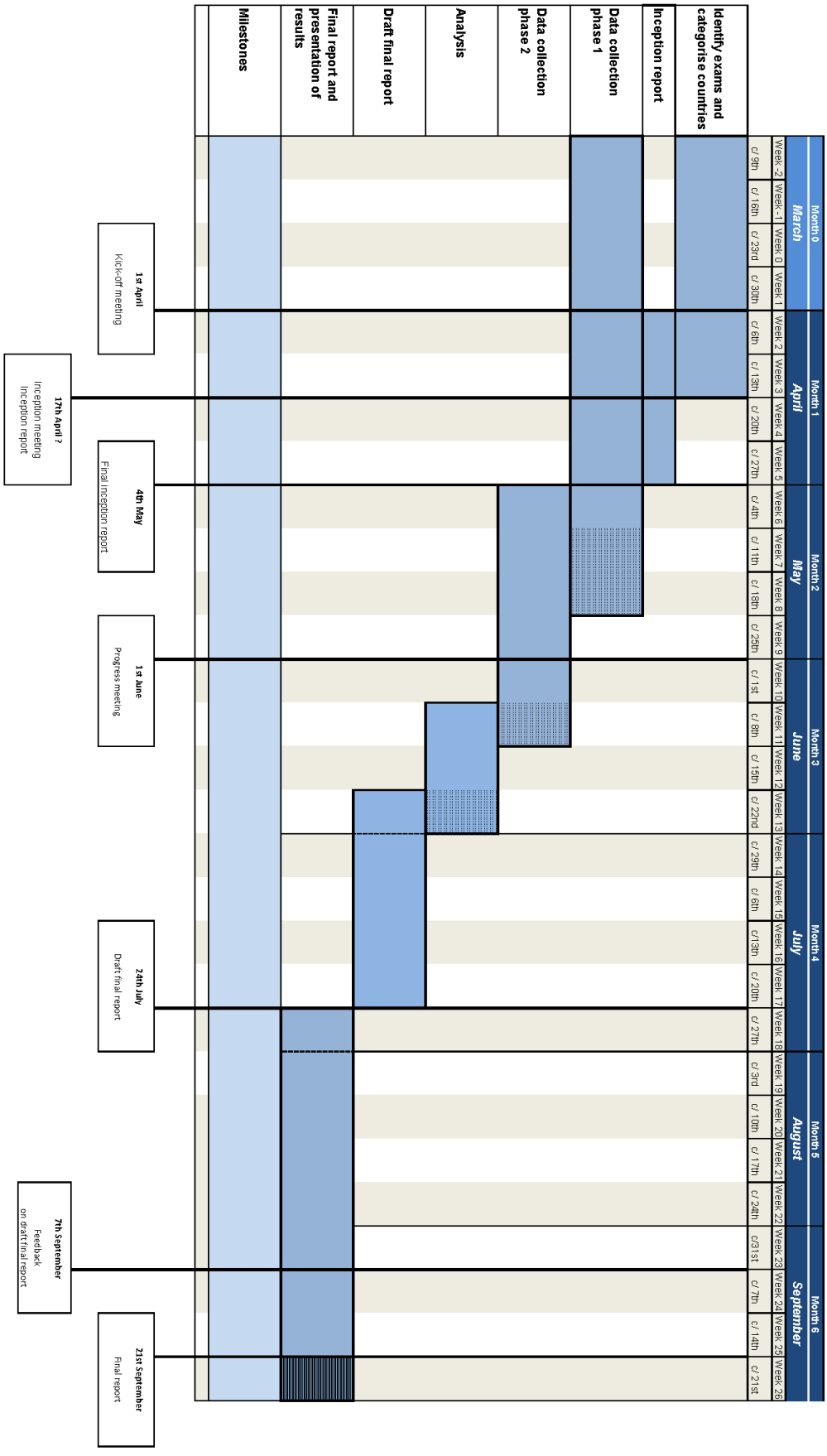
2.3.1 Contribution from EU Member States

In order to ensure that the results of this study are as accurate and transparent as possible, the European Commission facilitated the collaboration of the members of the Indicator Expert Group on Multilingualism (IEG). These members are all experts in language education and/or language assessment working for the Ministries of Education or National Statistical Offices in their respective jurisdictions.

After an initial meeting with the European Commission and the above-mentioned group of experts, the Project Team established direct contact with each of the members of the group to discuss in more detail the national language tests existing in each jurisdiction. The members' contribution was key to confirm the languages and tests chosen for each jurisdiction, and to provide any additional information regarding the exams (test papers, samples of performance, supporting documentation regarding the tests e.g., procedures for the creation and administration of exams, training materials for item writers and raters, national results, etc.).

2.3.2 Timeline

The above-mentioned Tender Specifications established the timeline for the project *a priori*, determining to a great measure the practical scope of the study. As observed in the timeline, while the length of the project was over 6 months, the majority of the work had to be completed in less than two months between the end of April and the end of June. This timeline was particularly problematic for the IEG members and representatives of the different jurisdictions, who in most cases have their busiest period in the year around May/June when they organise and run the national exams in their respective jurisdictions. This short timeframe and clash with national exam periods made the collection of relevant materials challenging, and in some cases explains why certain exams had to be excluded from the study.


Month 0
March
Week -2
Week -1
Week 0
Week 1
c/ 9th
c/ 16th
c/ 23rd
c/ 30th
Month 1
April
Week 2
Week 3
Week 4
Week 5
c/ 6th
c/ 13th
c/ 20th
c/ 27th
Month 2
May
Week 6
Week 7
Week 8
Week 9
c/ 4th
c/ 11th
c/ 18th
c/ 25th
Month 3
June
Week 10
Week 11
Week 12
Week 13
c/ 1st
c/ 8th
c/ 15th
c/ 22nd
Month 4
July
Week 14
Week 15
Week 16
Week 17
Week 18
c/ 29th
c/ 6th
c/13th
c/ 20th
c/ 27th
Month 5
August
Week 19
Week 20
Week 21
Week 22
Week 23
c/ 3rd
c/ 10th
c/ 17th
c/ 24th
c/31st
Month 6
September
Week 24
Week 25
Week 26
c/ 7th
c/ 14th
c/ 21st
Identify exams and categorise countries
Inception report
Data collection phase 1
Data collection phase 2
Analysis
Draft final report
Final report and presentation of results
Milestones
1st April
Kick-off meeting
17th April ?
Inception meeting
Inception report
4th May
Final inception report
1st June
Progress meeting
24th July
Draft final report
7th September
Feedback on draft final report
21st September
Final report

2.4 Structure of the report

As requested in the Tender Specifications, this final report contains 5 main sections addressing the main issues regarding the comparability of existing language examinations. An Introduction precedes these 5 main sections with some general notes about the background and practicalities of this study.

The 5 main parts of the report are outlined here as per the Tender Specifications (p.17), followed by the additional notes on each of the tasks to be accomplished in each section as specified also in the Tender Specifications (p.14).

Task 1: Assessment of comparability of the existing national language tests administered to secondary school students.

- Produce a critical yet constructive overview of comparability between different existing national or regional methods to assess language competences in Europe's secondary schools.

Task 2: Proposals for ex-post adjustment that can increase the comparability of existing results of language tests at national level.

- Identify and describe in detail proposals for measures and methodological procedures potentially needed to adjust for methodological differences in the results of existing national tests, in order to present the country aggregates in a coherent and meaningful European overview. This task directly concerns those jurisdictions that already have a national or regional system of language testing.

Task 3: Proposals for development work that could be undertaken at the national level to increase comparability of existing language tests.

- Identify and describe in detail proposals for development work that could be implemented by EU MS already having national or regional language tests, in order to increase the European comparability of their data. This task directly concerns those jurisdictions that already have a national or regional system of language testing.

Task 4: Proposals for development work that could be undertaken at national level by Member States not having a system for language testing and interested in developing one.

- Identify and describe in detail proposals for development work that could be implemented by EU MS not having implemented a national or regional language tests yet, with an approach that yields results comparable to other European jurisdictions.

Task 5: Comparative overview of existing country data on language testing

- Compile an overview of country data on language testing.

As stated in the Tender Specifications (p.18), while parts 1 to 4 include any foreign languages according to the criteria specified in page 5 of this report, part 5 will only include results for the first foreign language taught in each jurisdiction in addition to the main language(s) of instruction.

2.5 Project Team

In order to deliver this project, it was necessary to mobilise a wide range of experts in different fields. The project was governed by a Project Board and run by a dedicated Core Project Team that counted on the advice and support of a Network of Experts from Cambridge English Language Assessment's offices in Europe and the language partners in the SurveyLang consortium (Goethe-Institut, Centre international d'études pédagogiques, Universidad de Salamanca, and University per Stranieri di Perugia in collaboration with the University per Stranieri di Siena). Table 3 offers more details about the team members and their role in this project.

Table 3 Members of the project team.

Project team	Members
Project Board ▪ **Provide ongoing expert advice and quality assurance**	**Dr Nick Saville**, Chair of the Project Board **Dr Hanan Khalifa** **Dr Ardeshir Geranpayeh** **Tim Oates** **Nigel Pike** **Martin Robinson** **Stephen McKenna**
Core Project Team ▪ **Responsible for project management and delivery of outputs**	**Dr Neil Jones**, Project Director **Esther G. Eugenio**, Project Coordinator **Rosey Nelson**, Data Management Officer **Kasia Vazquez**, Project Assistant **Dr Michael Corrigan**, Analyst **Dr Joanne Venables**, Analyst

The Project Team also counted on the feedback and support of a Network of Experts from all over Europe, as well as with the help of a dedicated team of experts who contributed towards the project as content analysts and raters in the online ranking exercise. They are listed in Appendix 0 below. We would also like to thank Dr Chris Wheadon, founder of the *No More Marking* website, for his assistance in carrying out the Comparative Judgement exercise.

3 Outline of stages in the project

3.1 Sourcing information

3.1.1 Identifying exams

The first step in this project was to identify the exams that were going to be the object of study. This step was significantly facilitated by the work done towards the Eurydice Report *Languages in Secondary Education: An Overview of National Tests in Europe* (European Commission/EACEA/Eurydice 2015). Once the exams had been identified, in some cases it was necessary to seek clarification and/or confirmation from the relevant member of the Indicator Expert Group on Multilingualism or, in their absence, through the Network of Experts or other in-country contacts. This led to the identification of the 133 language examinations that would be included in this study from 33 different jurisdictions (28 EU Member States).

When more than one examination existed in a certain jurisdiction for the same ISCED level, preference was given to compulsory examinations, examinations taken by the largest number of students at that level, examinations at the end of that ISCED level, and examinations for which materials were readily available or easy to retrieve through the IEG members or the Network of Experts. Appendix 1 offers a detailed list of the exams included in this study, as well as the reasons why certain exams had to be excluded.

3.1.2 Categorising countries

As explicitly required in the Tender Specifications, the proposal suggested an initial categorisation of the 33 jurisdictions. After careful exploration of the data provided by the Eurydice Network, the categorisation suggested classifies jurisdictions according to the ISCED level at which they have language examinations, as shown in Table 4.

Table 4 Categorisation of jurisdictions according to their national language examination systems

Category	Number of jurisdictions
With ISCED 2:	27 jurisdictions
With ISCED 3:	30 jurisdictions
With ISCED 2 only:	3 jurisdiction
With ISCED 3 only:	6 jurisdictions
With both ISCED 2 and ISCED 3:	24 jurisdictions

3.1.3 Completion of source data

Once all the relevant exams had been identified, it was necessary to collect further information about these tests in order to undertake the comparability study. Important aspects could be extracted or inferred from the information provided by the Eurydice Network, and most of the remaining information could be obtained through a careful, expert analysis of the test papers that most jurisdictions make available on

their websites and that were collected and stored by the Core Project Team. Any missing documents for each jurisdiction (training materials for item writers and raters, sampling procedures when sampling of students occurs, test specifications, rating scales, quality assurance procedures for the creation and administration of tests, etc.) were requested and obtained in the majority of the cases through cooperation with the IEG members and, if necessary, through the Network of Experts, our staff at the Cambridge English Language Assessment offices in Europe and the ALTE partners in the relevant jurisdictions.

The online platform *Basecamp* was used as the main means of communication between the Core Project Team and the IEG members and country experts. This platform was particularly useful to store and monitor all previous emails exchanged with the IEG members or country experts in one place, especially considering the large amounts of data and emails that had to be handled simultaneously from 33 different jurisdictions in a very short period of time. For technical reasons, some experts were not able to join Basecamp, in which case traditional email was used.

The list of documentation requested from IEG members and experts included:

- samples of exam materials, preferably current exam papers (if unavailable, past exam papers)
- samples of performance, i.e. samples of students' Writing and Speaking
- statements of curricular objectives regarding language education
- test specifications
- rating criteria
- rating/assessment scales
- procedures for creation and administration of exams
- training materials for item writers and raters
- reports on statistical analyses
- sampling procedures (when sampling is used).

Test papers were collected and analysed for all the examinations included with the exception of Folkeskole Leaving Examination (Denmark, ISCED 2) and General Upper Secondary School Examination – STX (Denmark, ISCED 3), where these materials were confidential and not available for this study. In these cases, the analysis was conducted only on the basis of the descriptive information provided by the Eurydice Network regarding these examinations.

While sample or previous test papers seem to be publicly available online in most jurisdictions and relatively easy to find (with the exception of Finland and Denmark), some of the other documents requested proved more challenging to obtain. For example, hardly any jurisdictions were able to provide samples of performance or training materials used with item writers or with raters. Curriculum objectives and rating criteria were also difficult to obtain, as were usable data on national performance levels. **As a result of the deficiencies in the data, it was only possible to develop a partial understanding of most jurisdictions and of comparability between them.**

3.2 Content analysis of data

The content analysis was conducted in order to examine the extent to which test results were likely to be comparable. It was carried out with the help of an analysis tool in the form of an online survey, which was completed by 16 content analysts with proven experience and expertise in language assessment. They received specific training to complete this task, including familiarisation with the analysis tool and with all other sources of information necessary to answer the questions in the online survey. All of them also had an advanced level of English, and of at least one of the other main languages involved in the project (French, German, Spanish, Italian, Dutch and Swedish).

Content analysts were assigned each a number of jurisdictions and language examinations from the list in Appendix 1, taking into consideration their language skills and familiarity with national education systems in different jurisdictions. They made use of the available data (information provided by the Eurydice Network, example tests and other documents provided by jurisdictions) to attempt to complete the questionnaire. The full text of the content analysis tool is included in Appendix 2. They also had access to a specific *Basecamp* platform, which served as a repository for additional training materials and as a forum for communication between the Core Project Team and the content analysts.

Although the experts who completed the content analysis are leading experts in language assessment and most of them have extensive experience conducting similar analyses of language examinations, their work went through a process of spot checking to identify any potential clerical mistakes and misunderstandings. As part of the quality assurance of the project, 30% of the examinations were also analysed by a second expert, which ensured the consistency and reliability of the judgements made by the content analysts and allowed allowed for any discrepancies in their understanding of the task to be identified and addressed.

For further information about the content analysis see section 4.3.1, and for the results see 5.1 below.

3.3 Comparative Judgement exercise

The Comparative Judgement (CJ) exercise was conducted to provide a basis on which to compare the results of the different tests within the study. Initially, this technique was intended to be applied to samples of students' performance for Writing and Speaking, which would have offered an insightful and objective overview of proficiency levels demonstrated by students at each ISCED level. However, it proved to be extremely challenging to collect samples of performance within the timeframe and scope of the project. For this reason, and in order to demonstrate the potential of this technique for future comparability studies, this part of the study included samples of test tasks for Reading and Writing for the first foreign language in each jurisdiction. As observed in Table 1 above, the first foreign language in all jurisdictions is either English or French, with the exception of Luxembourg, where it is German.

Due to the use of tasks instead of samples of performance, the goal of the exercise shifted from establishing which of two samples of Writing or Speaking showed a higher level of language proficiency to determining which of two tasks was more difficult. The exercise was completed by 49 experts in language assessment and language education who received very specific instructions on how to conduct this exercise and the factors that they should bear in mind before making each of their judgements.

Particularly in the case of Reading tasks, they were asked not to base their judgements only on the basis of the difficulty of the texts but rather to consider the difficulty of the task as a whole, including the demands set by the items associated to the given text. However, experts doing this exercise had no additional information regarding the contextual conditions under which the exams were delivered, such as the length of time given to complete the tasks or the support offered to students to complete the tasks (e.g. use of dictionaries). In this study, therefore, such factors could not be taken into consideration when making the binary judgements.

The Reading and Writing tasks were selected and extracted from the test papers used for the content analysis. Due to the limited scope of this exercise and the main purpose being to show the potential of this technique for future comparability studies, only two tasks were selected from each included language examination for which Reading and Writing papers were available. When these papers included more than two tasks, by default the first and last tasks of each paper were extracted, assuming that difficulty would progress within the test paper from easier to more difficult. The tasks were then catalogued, labelled and anonymised, i.e. any references to the examinations and jurisdictions they came from were deleted to ensure objectivity when making the judgements. In some cases, the instructions or even the items appeared in a language different from the language being tested, which added an extra layer of difficulty to the exercise. When possible, a tentative translation of the instructions was provided in the extracted tasks to facilitate the work of the experts, who were only requested to have a good command of the target language being tested.

All the tasks were then uploaded into a website called *No more marking* (www.nomoremarking.com). This website has been designed to support an innovative approach to marking test papers which replaces subjective marking of papers in the traditional way (i.e. by applying a given assessment scale) with a ranking exercise of the students' performances based on repeated binary judgements of pairs of performances where the aim is just to determine which of the two is better. The website allows uploading all the samples of performance into the system, allocating a number of judgements to each expert, and then randomly showing experts pairs of papers and systematically collecting their judgements. Each pair of samples of performance is judged several times by different judges until the system considers to have collected enough evidence to reliably rank all the samples from the worst to the best. By feeding this into a Rasch model, and including in the exercise samples which have been previously assessed as representative of a certain CEFR level, it is possible to create a scale which shows the CEFR level demonstrated in each paper. For the reasons given above, in our study the samples of performance had to be replaced by Reading and Writing tasks, and the resulting scale provided an overview of task difficulty rather than students' performance in the target languages, as it would have been desirable.

For further information about the CJ exercise see section 4.3.2 below and for the results see section 5.2 below.

4 Concepts and approaches

4.1 The concept of comparability

We identify three major aspects to comparability:

1. The *construct* dimension, relating to the *validity* of the assessment: the conceptual framework which defines the language proficiency constructs which are tested and the interpretive measurement framework which is provided by the CEFR.
2. The *assessment* dimension, relating to the *reliability* of the assessment, which focuses on technical features of item design, method of test delivery, frequency of assessment, etc.
3. The *performance* dimension, relating to features of the candidature: their age and ISCED level, the language tested, etc.

In this study we have set out to implement comparison through these three hierarchically-ordered dimensions.

4.1.1 The construct dimension

The study explores three aspects:

- the purposes which language education is to address;
- the way in which the language skills of interest are defined;
- how progress in acquiring those skills is measured.

These substantive issues are addressed primarily in the qualitative study involving expert analysis of available documentation (section 4.3.1 below). They impact on comparability because jurisdictions may differ in how each of these aspects is conceived, prioritised, and implemented.

While respecting each jurisdiction's priorities, it is evident that an inclusive framework for discussion is still needed, and the CEFR is an obvious reference point. Linking to the CEFR has qualitative and quantitative (measurement) aspects which are addressed in the following sections.

4.1.2 The assessment dimension

The assessment dimension is where the constructs defined above are implemented in the form of test tasks or other forms of evaluation. Concerning comparability:

- Implementation may represent the intended constructs more or less validly, so that what is actually measured may not reflect intentions.
- The test may be more or less reliable. Lower reliability naturally impacts on comparability.

Note that comparability here includes comparability of different test sessions within the same jurisdiction. The issue of whether test versions vary in their level of challenge across sessions was raised in the country advisory group, but hardly any jurisdiction has provided data enabling this aspect of comparability to be examined.

4.1.3 The performance dimension

The performance dimension concerns evidence: to compare one jurisdiction with another data on performance are required, i.e.:

- responses to writing tasks, recordings of speaking tasks, annotated to show the marks awarded
- tables of test scores from objectively and subjectively-marked tests, showing the profile of achievement, and the interpretation of standards attributed to scores.

Concerning comparability, this is an area where relevant data has been hard to come by, but we will attempt to provide a model for how such data might be used in future comparability studies.

4.2 The CEFR as a framework for comparison

We adopt the Common European Framework of Reference as our comparative framework for examining construct and performance-related aspects. The text of the CEFR is available at: http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp. Several ancillary documents are also available from this weblink.

There are several reasons for adopting the CEFR:

4.2.1 As a familiar point of reference

The CEFR is widely referenced in Europe in relation to defining the goals of language education, professional training of teachers, curriculum development, and as a scale for reporting learning outcomes.

4.2.2 As a relevant model of learning

Given its multiple authorship, the CEFR speaks with several voices on the nature of language learning, but at its centre is the *action-oriented* model which sees language skill developing through motivated interaction within society. This essentially social-constructivist, socio-cognitive model has, we believe, general relevance to the language education goals of all jurisdictions. It is reflected in the Cambridge English Language Assessment approach to assessment, and also in the questions posed in the expert analysis of tests (4.3.1 below).

4.2.3 As a measurement construct

We may identify two distinct aspects to the CEFR. As presented above, it provides a detailed discussion of how languages may be taught and learned as tools for communication. It also presents a framework of levels, which sets out to enable a broad comparison of language learning programmes and purposes. Arguably it is as a framework of levels that the CEFR is best understood and mostly referred to.

The set of CEFR level descriptors A1 to C2, which have been widely adopted within Europe and beyond, is based on tables of can-do statements. These were compiled through an empirical study, using item response theory. More importantly, the CEFR levels have been adopted and developed by several examination bodies, primarily Cambridge English Language Assessment and ALTE, as a scale for reporting levels of achievement in language tests. Thus there are excellent models available for jurisdictions who wish to construct their own assessments and link them to the levels of the CEFR.

As the reporting scale adopted for the first European Survey on Language Competences (European Commission 2012) the CEFR levels were used to provide benchmark measures of language achievement in European schools. It is highly desirable that the study reported here should be anchored to the same scale, and going forward, we should see projects focused on CEFR levels as potentially useful elements in a movement to bring national or regional language assessments into alignment.

Appendix 6 outlines the six CEFR levels A1 to C2, taken from the CEFR.

4.3 The approach: qualitative and quantitative

4.3.1 The qualitative analysis focus: expert analysis of tests

Over the years several groups have attempted to develop descriptions of language use and of the contextual features which determine the difficulty of tasks, and consequently the nature of progression in learning. The CEFR itself provides a large number of scales for different aspects of language use. Critical appraisal of these has led others, such as the Dutch Constructs Group (Alderson et al 2004, 2006), ALTE (ALTE 2005) and Cambridge (the English Profile, e.g. Green, 2012) to develop a range of instruments for analysing the content of tests. Some of these are available within the Council of Europe's CEFR 'toolkit', which includes two major documents: a *Manual for relating Language Examinations to the Common European Framework of Reference for Languages (CEFR)* (2009), with a technical Reference Supplement, and the *Manual for Language Test Development and Examining* produced by ALTE on behalf of the Language Policy Unit of the Council of Europe in 2005.

The qualitative expert analysis was organised through a questionnaire which 16 highly-competent experts in language assessment completed online. Additional reference material was provided to experts to assist in completing the questionnaire.

The questionnaire's broad sections are:

1. Introduction
2. The exam/ test: high level description
- Design and purpose
3. Goals of language education
4. Speaking:
- Rating
- Speaking: tasks
5. Writing:
- Rating
- Writing: task input/ prompt
6. Reading

7. Listening

8. Structural competence

The sections dealing with the four skills require experts to apply the questionnaire to each task testing that skill (up to a maximum). The full text of the questionnaire is included as Appendix 0 below, and section 5.1 below provides detailed outcomes of the analysis.

The development of the questionnaire has benefited from reference to recent projects undertaken by Cambridge English Language Assessment for ministries of education and educational initiatives in a number of jurisdictions. An interesting feature is the attempt to build levels of cognition into descriptions of test tasks, in addition to the more common behavioural descriptors. The success of this is discussed in section 5.3 below, where we combine the outcomes of the qualitative and quantitative analyses.

4.3.2 The quantitative analysis focus: Comparative Judgement (CJ)

Comparative judgement is an approach to constructing measurement scales using item response theory (IRT). It uses an extension of the Rasch model which allows tasks from different tests to be calibrated onto a single difficulty scale, because the judges' responses constitute a single, linked dataset. Comparative judgement thus enables us to compare exam tasks chosen from a range of different exams and, most importantly, samples of students' performance for Writing and Speaking, in a way which is much more practical than organising a data collection requiring students to respond to a specially-assembled set of inter-related test tasks.

The CJ method employed in this study uses the Bradley-Terry model (Appendix 5 below). Comparative Judgement is effective because it asks the rater to do something fairly simple: to compare two exemplars and say which is better, or, in the case of the test tasks compared in this study, harder. It is a relative judgement, not an absolute one, as is the case of marking candidates' test performances. Comparative Judgement seems to be gaining ground as an effective alternative to marking, or as an approach to critiquing it. As an example, a significant study by the exams watchdog Ofqual in the UK shows the differences in standards for mathematics invoked by a range of exam providers. That study was conducted on the *No More Marking* website, and an interesting interactive report is available there. This is the site we have also used for the Comparative Judgement exercise described here.

The relative simplicity of Comparative Judgement makes it accessible to a wider range of stakeholders in education; and the more judges are involved, of course, the better the precision of estimation. It remains important that raters should share an understanding of the bases of comparison, so that some training or guidance is certainly desirable. For this project, expert raters were specifically recruited by Cambridge English Language Assessment to participate in the comparative study, some of whom also participated as experts in the qualitative content analysis. Raters worked online from home and were given detailed instructions on how to complete the task and what aspects they had to take into consideration as the basis for their judgements. Analysis of response patterns and agreement between groups in fact showed relatively insignificant differences in performance between the different raters.

5 Task 1: Assessment of comparability of existing national language tests

5.1 Qualitative analysis: Expert descriptive analysis of tests

The aim of the qualitative analysis was to examine the extent to which the language tests under investigation were likely to be comparable. Differences in the construct (what is being tested), the populations for which the tests are intended, the basis for interpretations of test results and characteristics of the measurement (features which add diversity and therefore decrease comparability) were all considered.

5.1.1 Method

The data comprised information on the language examinations in Appendix 1 and for the languages identified in Table 1. The collection of the data used in this section is described in section 3.2, and the content analysis tool is described in section 3.3.1 above. For the full data collection instrument, please see Appendix 2.

After the data was extracted from the online survey provider on which the content analysis tool was built, it was partially recoded and collated for the descriptive analysis. The aim of the recoding was to ensure constructed responses were compared in a meaningful way, despite various differences in expression for essentially the same aspect. For example, total exam duration was sometimes expressed in minutes and sometimes in hours and minutes. This information was recoded so that all responses were expressed in minutes. Collation was also necessary, so that each case in the data file represented responses for a specific test, rather than confounding all tests provided by a single jurisdiction at one ISCED level.

For the descriptive analysis of this data, charts and tables were produced for each item using Microsoft Excel 2010. For all questions, since test comparability was the topic of interest, it was important to ensure that tests were compared on the basis of what was expected to be similar. For analysis of questions concerning the test construct, therefore, each skill was examined separately. Table 5 Skills tested at ISCED 2 level in each jurisdiction.and Table 6 show the language skills tested in each jurisdiction and ISCED level, which should be borne in mind when interpreting the results of the analysis of each construct.

Table 5 Skills tested at ISCED 2 level in each jurisdiction.

Jurisdiction	Reading	Writing	Listening	Speaking	Language Use
Austria	✓	✓	✓	✓	
Belgium FR	✓	✓	✓	✓	
Belgium GE	✓	✓	✓	✓	
Belgium NL	✓	✓	✓		
Bulgaria	✓	✓	✓		
Czech Rep			✓		

Denmark	✓	✓	✓	✓	
France	✓	✓	✓		
Germany	✓		✓		
	✓		✓		
Hungary	✓		✓		
Ireland	✓	✓	✓	✓	
Latvia	✓	✓	✓	✓	
Lithuania	✓	✓	✓		
Luxembourg	✓	✓	✓		
	✓	✓			
	✓				
Malta	✓	✓	✓	✓	
Netherlands	✓	✓	✓		
Poland	✓	✓	✓		✓
Portugal	✓	✓	✓	✓	
Romania	✓	✓			
Slovenia	✓	✓	✓		
Spain - Navarre	✓	✓	✓		
Spain - Catalonia	✓	✓	✓		
Sweden	✓	✓	✓	✓	
UK England	✓	✓	✓	✓	
UK Northern Ireland	✓	✓	✓	✓	
UK Scotland	✓	✓	✓	✓	
UK Wales	✓	✓	✓	✓	

Table 6 Skills tested at ISCED 3 level in each jurisdiction.

Jurisdiction	Reading	Writing	Listening	Speaking	Language Use
Austria	✓	✓	✓	✓	
Belgium GE	✓	✓	✓	✓	
Belgium NL			✓	✓	
Bulgaria	✓	✓	✓		
Croatia	✓	✓	✓		
Cyprus	✓	✓	✓		
Czech Rep	✓		✓		
Denmark		✓	✓	✓	
		✓	✓	✓	
Estonia	✓	✓	✓	✓	
Finland	✓	✓	✓		
France	✓	✓	✓	✓	
Greece	✓	✓			
Hungary	✓	✓	✓	✓	
Ireland	✓	✓	✓	✓	
Italy	✓	✓			
Latvia	✓	✓	✓	✓	
Lithuania	✓	✓	✓		
				✓	
Malta	✓	✓	✓	✓	
Netherlands	✓				
Netherlands	✓				
Poland	✓	✓	✓	✓	✓
Portugal	✓	✓			
Romania	✓	✓	✓	✓	
Slovakia	✓	✓	✓		

Slovenia	✓	✓	✓	✓	
	✓	✓		✓	
Sweden	✓	✓	✓	✓	
UK England	✓	✓	✓	✓	
UK Northern Ireland	✓	✓	✓	✓	
UK Scotland	✓	✓	✓	✓	
UK Wales	✓	✓	✓	✓	

A more specific consideration of the test construct and test materials (i.e. what is being tested) was also important, and the stated CEFR levels of the test were used for this purpose. CEFR levels are useful for this purpose, as each successive level implies a variation in construct. This is because each CEFR level is nested in the one above. A B1 level learner would, therefore, be expected to do everything an A2 level candidate could do, as well as something additional or better. As a consequence, data was matched by CEFR level, so that, for each skill, only those tests which attempted to measure at the same level were compared. Where CEFR level was not given, the data was excluded from this analysis, as inclusion would have led to confusion over what was being compared. Where tests were recorded as targeting a range of CEFR levels, the test was compared to others on the basis of the highest level it tested. This relates to the logic of the nested structure of CEFR levels mentioned above. The highest level would provide the most complete description of what was tested, as preceding levels would also be represented. The findings of this analysis are available in section 5.1.2.

Test materials were obtained from test providers and examined by the expert content analysts. This enabled a comparison of the level stated in the Eurydice data and the estimated CEFR levels according to the expert analysts. as an approach to verifying the stated CEFR target. Poor targeting would indicate lack of comparability of test results reported in terms of the CEFR, since it implies a mismatch between the things a candidate can do and those which the reported CEFR level implies. Findings are reported in section 5.1.3, which focusses on the interpretation of test results.

When analysing questions not specifically related to the construct or CEFR level, the data were grouped by ISCED level or as a single group. This was because European Commission/EACEA/Eurydice (2015) found that ISCED 3 tended to represent higher stakes tests, and therefore it was thought likely that differences may have occurred in terms of some procedures and other characteristics, such as the thoroughness with which alignment to the CEFR was established (see section 5.1.3). In other cases, if it was thought likely that an institutional influence would be stronger, the data were analysed in one group – for example, the recruitment and training of staff (see section 5.1.3). Variability discovered in tests in this analysis would therefore suggest lack of comparability between tests in terms of interpretations of results (section 5.1.3), populations (section 5.1.4) and measurement characteristics (section 5.1.5), which deals with aspects which might affect the accuracy and reliability of measurement.

5.1.2 Comparability of constructs – to what extent the exams measure the same abilities

Although all tests considered in this study test foreign language ability, the definition of ability, target language apart, may vary considerably. This section will investigate some of the key features considered when defining language ability for language tests. Following Weir (2005), each skill was examined separately because many of its most important characteristics are unique. Speaking and Writing share some similarities, however, as tests are usually based on eliciting a sample of performance from the candidate. Reading and Listening are also similar to each other, as such tests usually involve an interaction between the candidate and texts presented to them.

Tests were compared according to their CEFR level in order to account for differences in the test construct due to target ability level. This was felt important, as for most key parameters, harder tests would be expected to be different from those at lower levels (Geranpayeh & Taylor, 2013; Khalifa & Weir, 2009; Shaw & Weir, 2007; Taylor, 2011).

Speaking

The ability to interact with others offers challenges that are clearly not present in monologues (Council of Europe, 2001). These include, for example, the need to understand the rules of turn-taking and other such mechanisms important for sustaining the interaction. An interaction with an examiner can be expected to be different from that with a fellow test-taker, primarily because of the difference in social relationships it implies. This, in turn, would mean different politeness rules and other such requirements.

Figure 1 shows that the *forms of interaction* in the tests studied vary for each CEFR level. Of the tests studied, all were reported as having no more than two different interaction patterns, with many having only one. Such a finding limits comparability across tests, because, as explained above, the nature of what is being tested is different in each case.

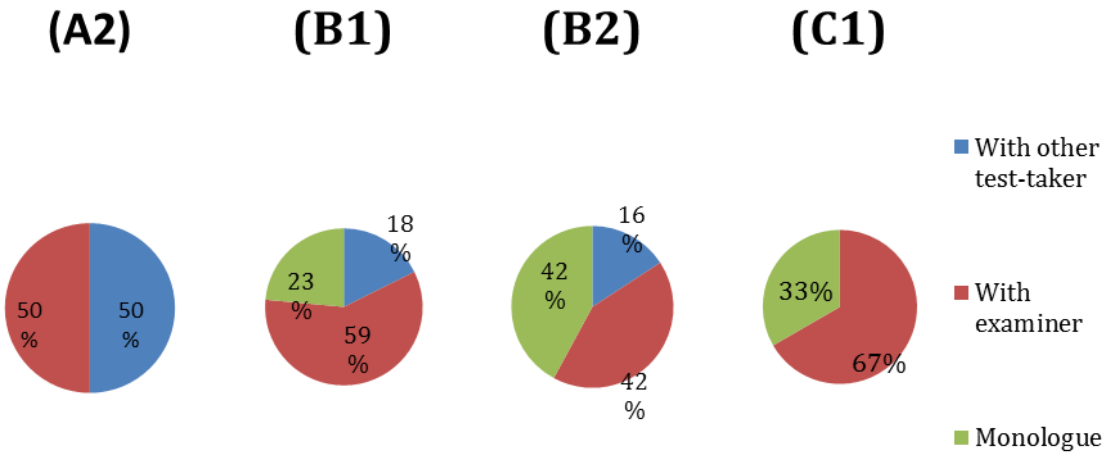


Figure 1 Interaction patterns averaged across tests

Figure 1 shows that at A2 level there is on average a balance between interaction with the examiner and with another test-taker. It is encouraging to see that peer-to-peer interaction is quite widely adopted at this level.

At B1 there is on average considerably less peer interaction and monologue is introduced, suggesting that the focus on basic communicative interaction at A2 is to an extent replaced by a more penetrating evaluation of the individual learner.

At B2 there is on average an equal amount of monologue and examiner interaction, again suggesting more formalised and perhaps rehearsed performance.

At C1 there is an exclusive use of monologue and examiner interaction.

It is interesting to observe that interaction between peers is generally deemed unsuitable for testing Speaking at higher levels.

The *goals* of any communicative task, whether in a test, or in a non-test situation, affect the way it is approached and conducted by the language learner. Linguistic features of what is said, as well as the structure of the performance and other aspects are affected. The communicative purposes in tests include in this study include:

- *conative*, i.e. arguing or persuading;
- *emotive*, or expressing a reaction to something;
- *phatic*, or simple social interaction;
- *referential*, i.e. telling or recounting something

Figure 2 shows a variety of purposes across levels, with, as would be expected, higher levels tending towards a greater range of purposes, as language use at these levels is more sophisticated. Most tests were found to test less than three communicative purposes at all levels when their tasks were examined. The fact that not all purposes were tested and that at higher levels, three or four purposes were used, suggests a greater lack of comparability at higher levels than lower levels.

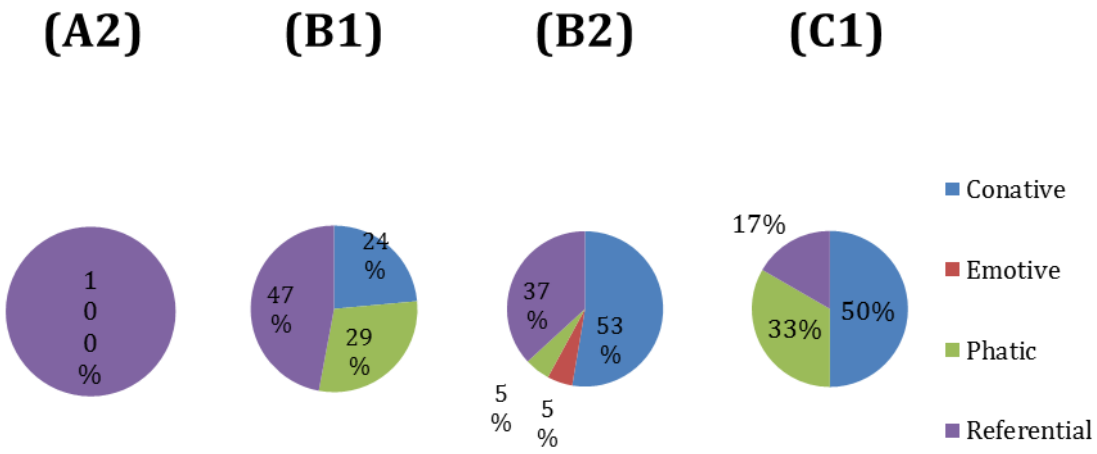


Figure 2 Communicative purpose of response averaged across tests

At A2 level, according to the data averaged across tests, it would appear that all interaction is referential - an unexpected finding, given the large proportion of peer-to-peer interaction at this level.

At B1 level phatic and conactive purposes appear, suggesting that the student is additionally expected to maintain an interaction and to argue a case.

At B2 level there seems to be a considerable stress on arguing a case. which perhaps reflects the high level of monologic presentation

At C1 this stress is maintained. The importance of phatic communication at this level is difficult to explain.

In summary, these communicative purposes do not offer a very clear picture of progression across the CEFR levels.

Domains are described by the CEFR as 'spheres of action or areas of concern… in which social life is organised' (Council of Europe, 2001:45). As such, they represent the context of language use. Their effect may be important, as they impact on acceptable forms of speech. For example, the educational domain would imply interactions with a teacher and this, in turn, would imply using language which appropriate to the social relationship between learner and teacher. In the personal domain, since different relationships prevail, different language might be appropriate. The domain would also influences the types of communicative task called for, likely topics of discussion. and so on.

The domains implied by the Speaking test materials examined are presented in Figure 3. As might be expected, the personal domain is more prevalent than at higher levels. By contrast, at higher levels, the public domain is more important. This notwithstanding, most tests did not represent topics from more than one or two domains, which implies some lack of comparability.

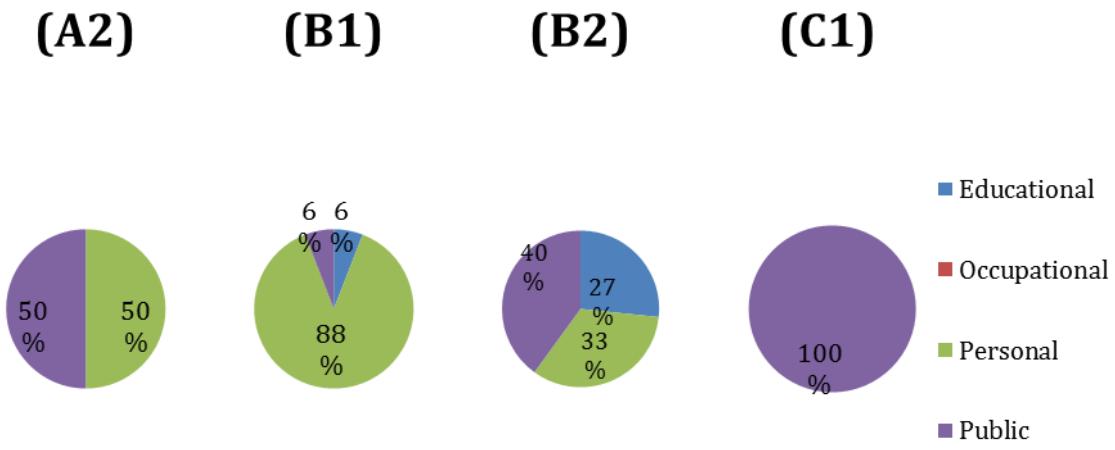


Figure 3 Domain of response averaged across tests

At A2 level there is an equal focus on personal and public domains: talking about oneself and relating to simple tasks in the world.

At B1 level the personal domain would appear to be dominant .

At B2 there is a significant focus on the educational domain.

At C1 there appears to be a total focus on the public domain, which is difficult to explain.

Interestingly, there is no use at any level of the occupational domain.

Rating criteria are of considerable importance to the definition of the test construct in performance (Speaking and Writing) tests. This is because it directs what raters are to evaluate and so mediates between the performance and the test result. In other words, whatever abilities the candidate displays in his or her performance, only those mentioned in the rating criteria have a direct influence on the test result.

Figure 4 shows the rating criteria used at different levels. These findings suggest a core of comparability between tests, with some differences. A future study might investigate how rating descriptors, training and other practices used to guide the use of these criteria compare. This might reveal further issues of comparability between tests.

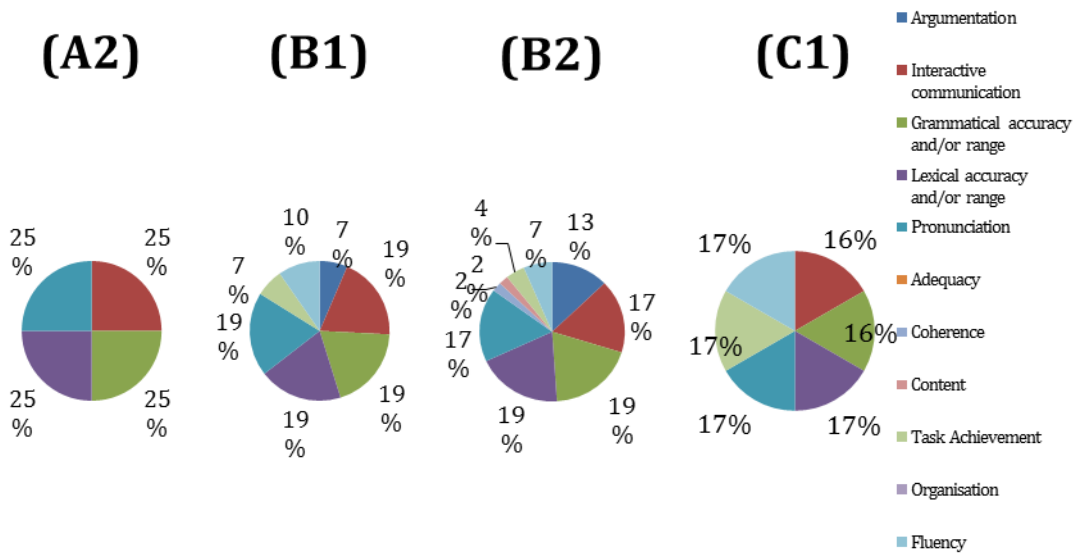


Figure 4 Rating criteria averaged across tests

At A2 level the criteria *Pronunciation*, *Interactive Communication*, *Grammatical Accuracy/Range* and *Lexical Accuracy/Range* suggest a broad concern with the basic aspects of Speaking, including its communicative effectiveness. These criteria continue to be used across all levels, with higher levels characterised additionally by *task achievement* and *fluency*, which indicate higher quality of performance.

Writing

Communicative purpose and *domain* are explained above under 'Speaking'. Their interpretation for Writing is identical. Figure 5 and Figure 6 indicate a similar level of variability across levels as for Speaking, and therefore, similar challenges for comparability.

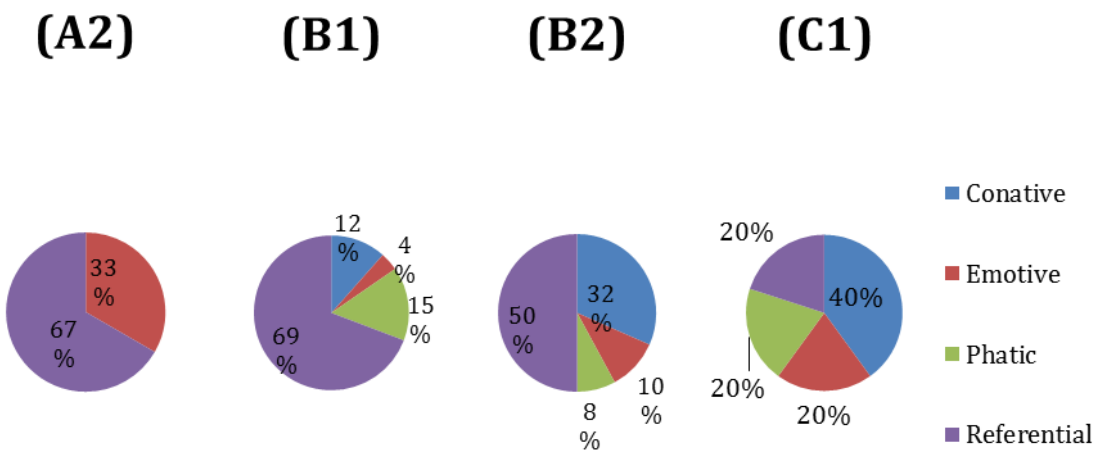


Figure 5 Communicative purpose of response

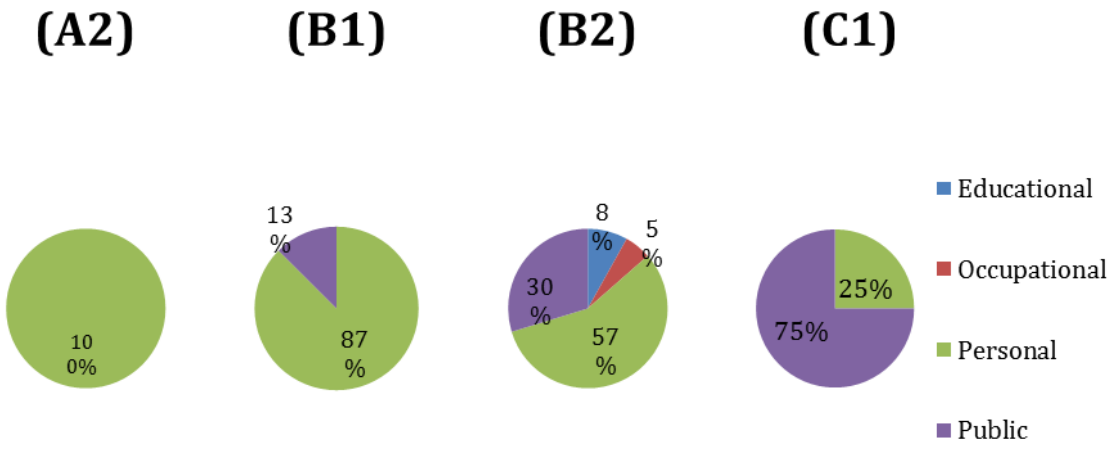


Figure 6 Domain of response

The expected *register* of response is an important aspect which is implied by the task. Formal register is usually marked linguistically and therefore requires learners to know the difference between formal and informal expression in the target language. As Figure 7 shows, the formal register becomes more important as the level increases, and the informal register declines in importance. In almost all cases, the test examined required candidates to produce a response in only one register, providing a greater challenge to comparability at B2 and C1.

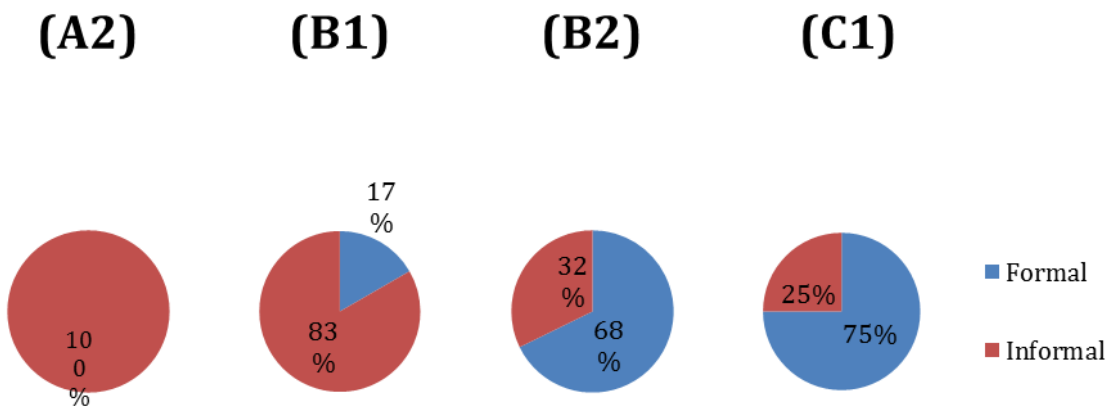


Figure 7 Expected register of response

As with Speaking, *rating criteria* have an important role in defining the construct which is represented in the test results. Figure 8 shows that three criteria were found at all levels: *Content*, *Grammatical Accuracy / Range* and *Lexical Accuracy / Range*. . As with the Speaking criteria, these criteria can be considered as core. Outside of this, there is some variation, which makes comparability more difficult.

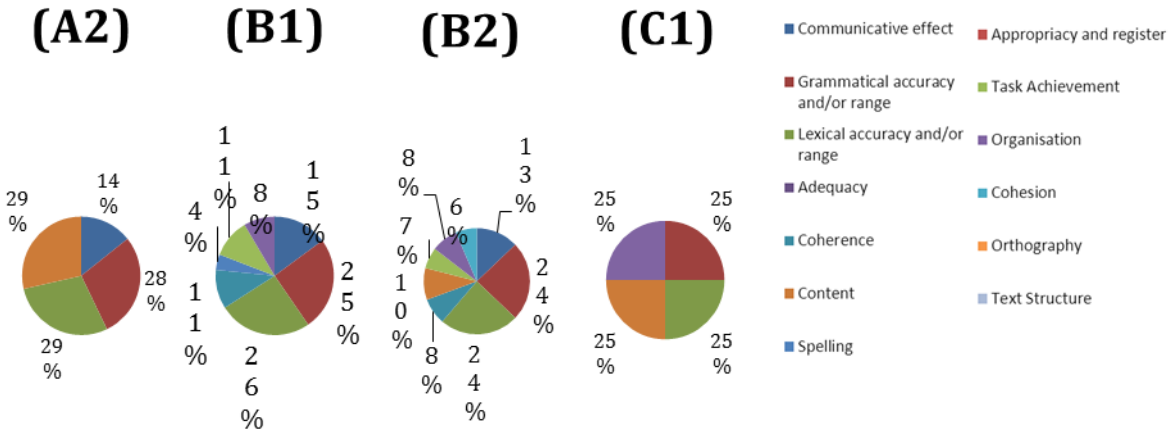


Figure 8 Rating criteria

Reading

For Reading, '*domain*' refers to that of the reading passage(s). Its meaning and interpretation is, identical to its use elsewhere in this report. The public domain is the most popular at all levels, as Figure 9 shows. The use of texts from the personal domain decreases at higher levels, as might be expected. In most cases, texts were

taken from one or two of these domains, suggesting greater lack of comparability at higher levels, where a greater range of domains were tapped.

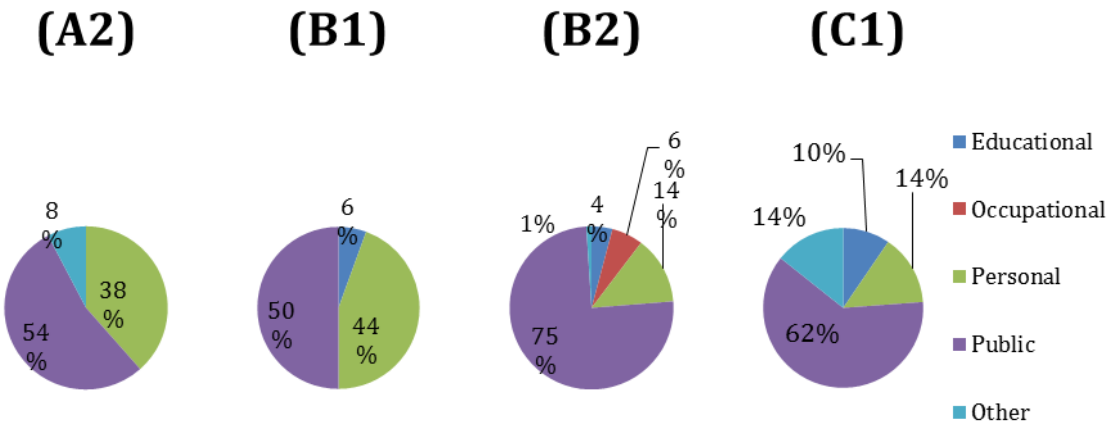


Figure 9 Topic domain averaged across tests

Different *types of reading* may be required by different tasks. According to Khalifa and Weir (2009), Reading may involve the processing of each word of the text so that individual units of meaning (propositions) are understood. On a global level, this encompasses large stretches of text and possibly the entire text. Careful local reading refers to small segments of text, where there is no need to make connections between propositions. Expeditious reading does not require the careful processing of every word. In non-test situations, this type of reading is evident when searching a newspaper for a story on a particular topic, or scanning the story, once found, for specific information within it. In everyday reading tasks, efficient readers use all types of reading to achieve their goals.

In Figure 10, careful reading, whether local or global, was found at all levels. Expeditious reading was tested at levels B1 to C1 only. Generally speaking, fewer than three types of reading were found in any one test, thereby increasing lack of comparability at higher levels where the range of reading types was greatest.

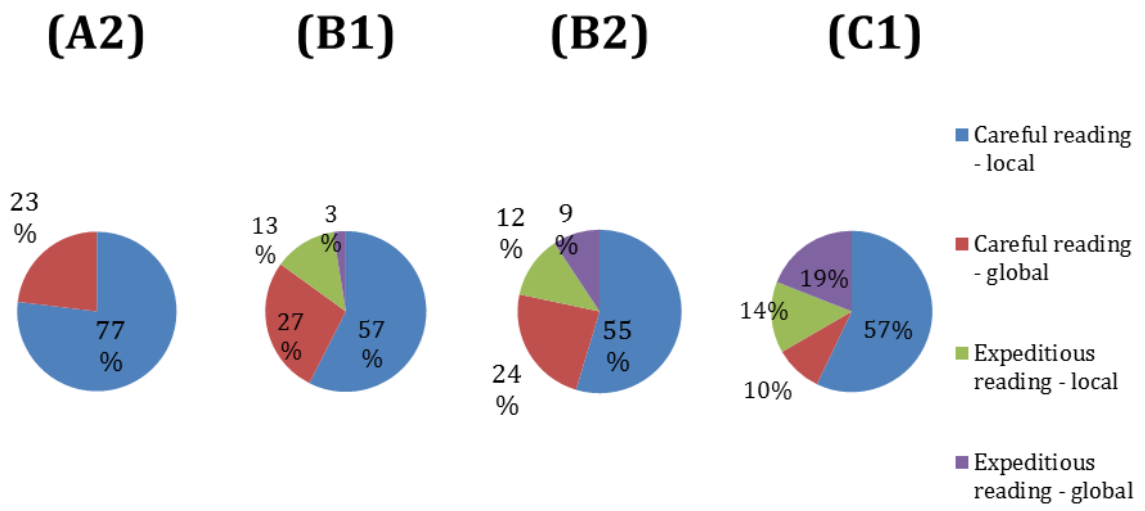


Figure 10 Type of Reading, averaged across tests

At A2 level there is a focus on careful local reading, as expected. What is striking is how dominant this type of reading remains across the levels. In the construct proposed by Khalifa and Weir the highest form of reading is expeditious global reading; which is indeed most evident at C1 level.

Examining the cognitive process of Reading can help to identify the sources of difficulty within a test task. Khalifa & Weir (2009) provide a model of the cognitive process of Reading which deconstructs the process eight stages. These stages are summarised in Table 7.

Table 7 Stages in cognitive processing model of Reading (Khalifa & Weir, 2009)

	Stage	Gloss
1	word recognition	creating a phonological representation of the word from the graphical representation of the word
2	lexical access	matching the word to a semantic representation stored in the mental lexicon
3	syntactic parsing	determining the syntactic role of each word (part of speech) in the sentence
4	establishing propositional meaning	identifying propositions and attaching a semantic meaning to each
5	inferencing	inferring information which is not explicitly given in the text
6	building a mental model	constructing a mental representation of the situation described by the text, independent of the words used to describe it in the text
7	creating a text level representation	developing a mental model for the entire text, or large parts of it
8	creating an intertextual representation	developing a mental model for multiple texts

In order for a reader to complete later stages described by Khalifa & Weir's (2009) model, it is first necessary to process text through the earlier stages. However, the model does not imply a rigid sequential progression through the stages: readers will jump backwards and forwards as they work through the text. The later stages, however, are considered to be more difficult than the earlier stages for several reasons. First they involve greater use of mental resources (more information must be held in the short term memory) and second, since later stages cannot be completed without earlier ones, the difficulties of all preceding stages influence the chances of success in any later stage. For these reasons, the highest stage of cognitive processing required for an item is expected to depend on the target level of the exam.

As Figure 11 shows, in line with expectations, there was a progression across levels, with later processing stages being found at higher levels. As there were a limited number of C1 tests, the range of processes tapped in this chart was curtailed. Due to the limited scope of the current research, examination conducted by expert analysts was at task level, such that the highest level of processing demanded by a single task was recorded. Greater variation might have been found at item level but such an investigation was beyond the scope of the current study. This may partly explain why, most tests were found to have a range of one, two or three levels which determined to be the highest reached. This also suggests a lack of comparability between tests, as, at B1 and B2, seven distinct cognitive levels appear on the chart.

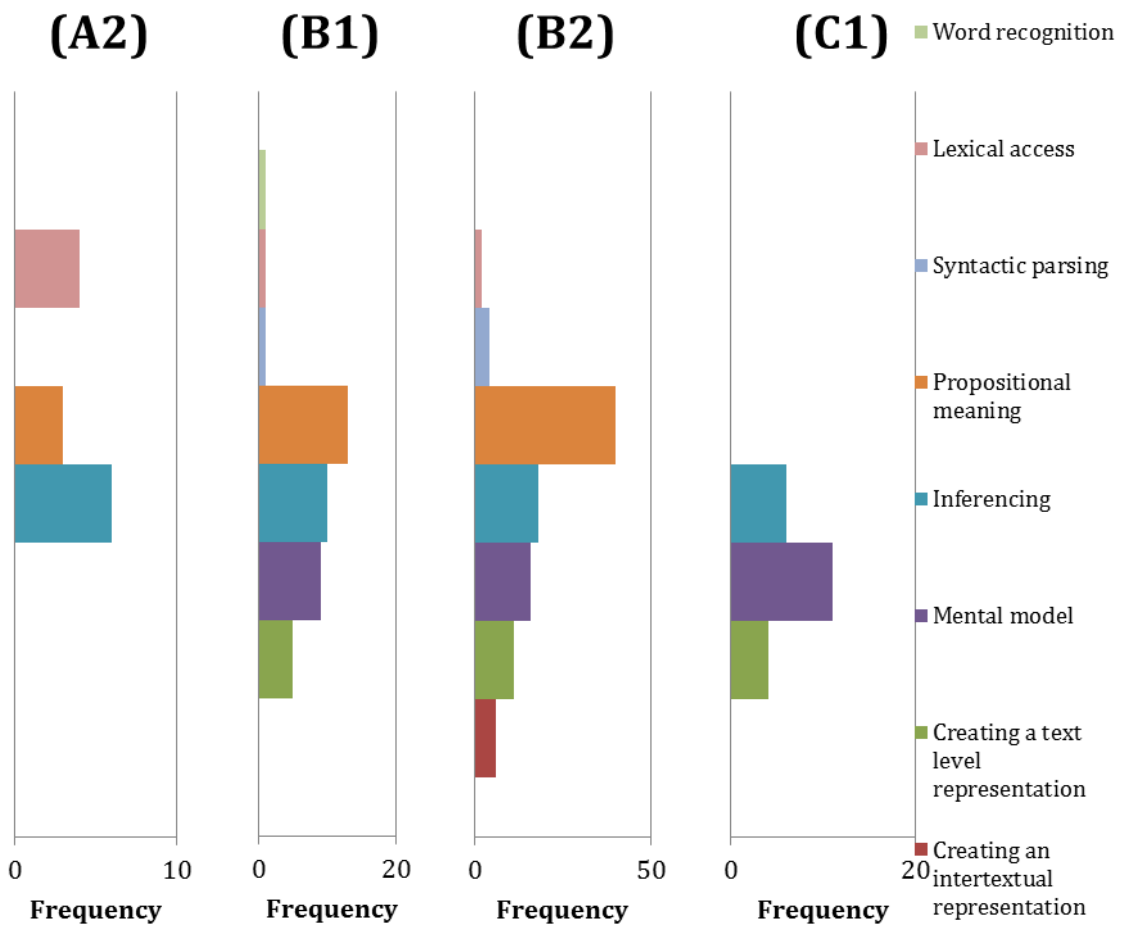


Figure 11 Highest level of cognitive processing (Reading)

A2 level is characterised by lower-level functions - understanding the meaning of words, and establishing propositional meaning (although some inferencing is also required).

At B1 level the highest level is *creating a text-level representation* i.e. a good understanding of the text.

At B2 level there is evidence of *intertextual representation*, i.e.understanding the text or texts within a wider context

At C1 level the highest function is at least *inferencing*; the narrower upper range reflects the limited number of examples.

Listening

As with Reading, Writing and Speaking, *domain* is to be interpreted in the way. Figure 12 shows a similar progression from personal to public across the levels to that of the other skills, with other domains playing little part. Most tests drew texts from no more than one domain at each level, and this implies some lack of comparability, given the range of levels found in all tests.

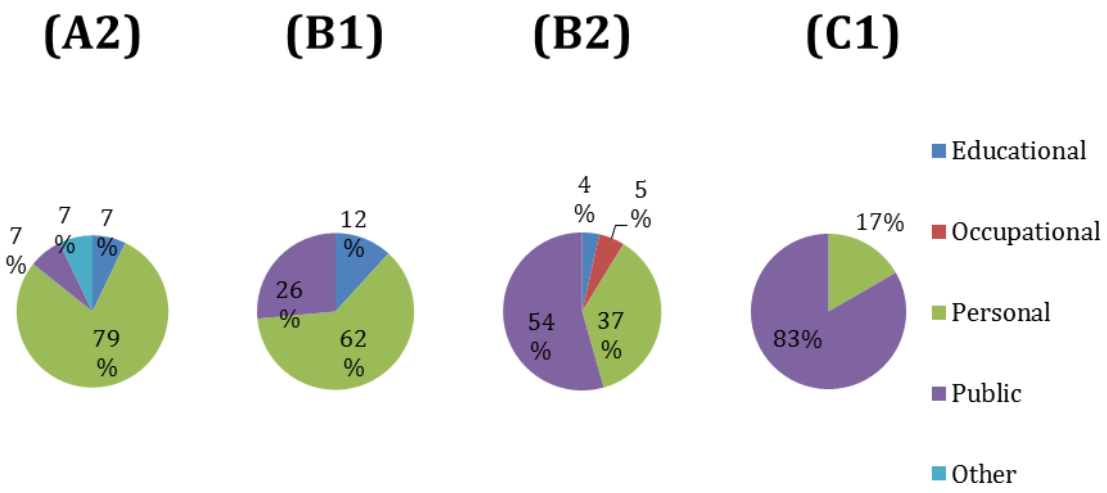


Figure 12 Listening topic domain averaged across tests

Types of listening are analogous to types of Reading, although, since listening occurs in real time, mental resources come under greater pressure and longer stretches of text become increasingly difficult to process. Careful local listening is an attempt to decode every word in short segments of the listening text. Careful global listening involves a detailed decoding of the entire text. Expeditious listening requires, for example, listening for information on a particular topic, listening for specific information or the gist of an entire text.

In Figure 13, in a similar fashion to the types of reading examined, careful local listening was the most widespread listening type. Careful global listening was also common, with relatively little expeditious listening required, except at B2. As most tests required only one type of listening, these findings suggest some lack of comparability.

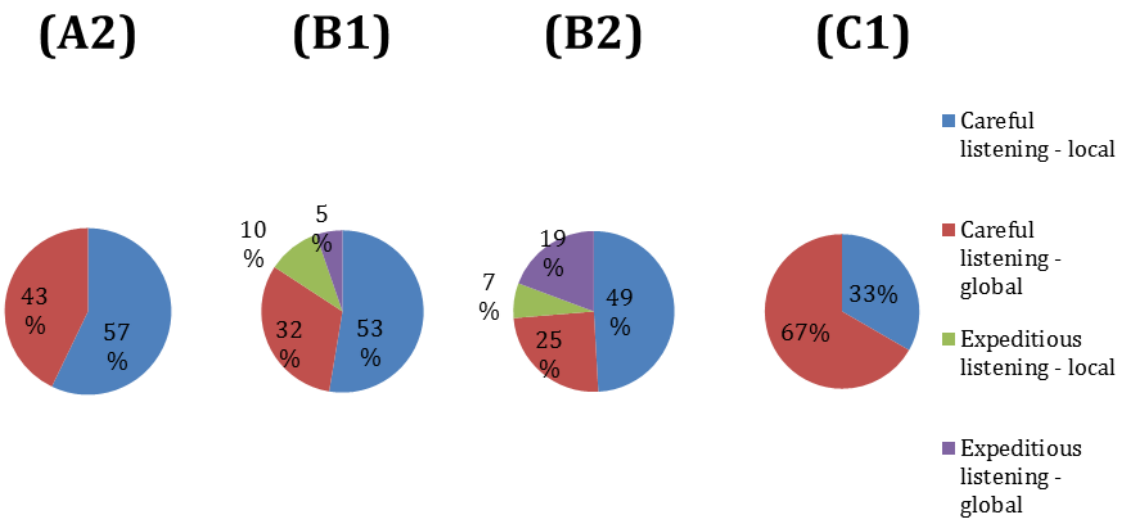


Figure 13 Type of Listening averaged across tests

A cognitive processing model has been created for Listening (Field, 2013) and bears many similarities to that of Reading (see above). The key stages of the model used in the current study, based on Field's (2013) model, are summarised in Table 8.

Table 8 Stages in cognitive processing model of Listening (Field, 2013)

	Stage	Gloss
1	input decoding	representation of speech signal so that it conforms to the phonological system of the target language
2	lexical access	matching the phonological representation to a semantic representation stored in the mental lexicon
3	syntactic parsing	determining the syntactic role of each word (part of speech) in the sentence
4	establishing propositional meaning	identifying propositions and attaching a semantic meaning to each
5	inferencing	inferring information which is not explicitly given in the text
6	creating a mental model	constructing a mental representation of the situation described by the text, independent of the words used to describe it in the text
7	creating a discourse representation	developing a mental model for the entire text, or large parts of it
8	creating an intertextual representation	developing a mental model for multiple texts

As can be seen in Figure 14, in line with expectations, and in a similar fashion to the findings for Reading (see above), a progression to higher cognitive levels is visible as the CEFR level increases. Unlike the cognitive processing requirements for Reading, however, there appears to be a ceiling effect, with neither of the two highest levels found to be required. This is perhaps because of the challenges in creating tasks which require the higher stages. At B1 and B2, five or six distinct levels were detected, whereas most tests only tapped two or three of them. This suggests significant lack of comparability within the Listening construct.

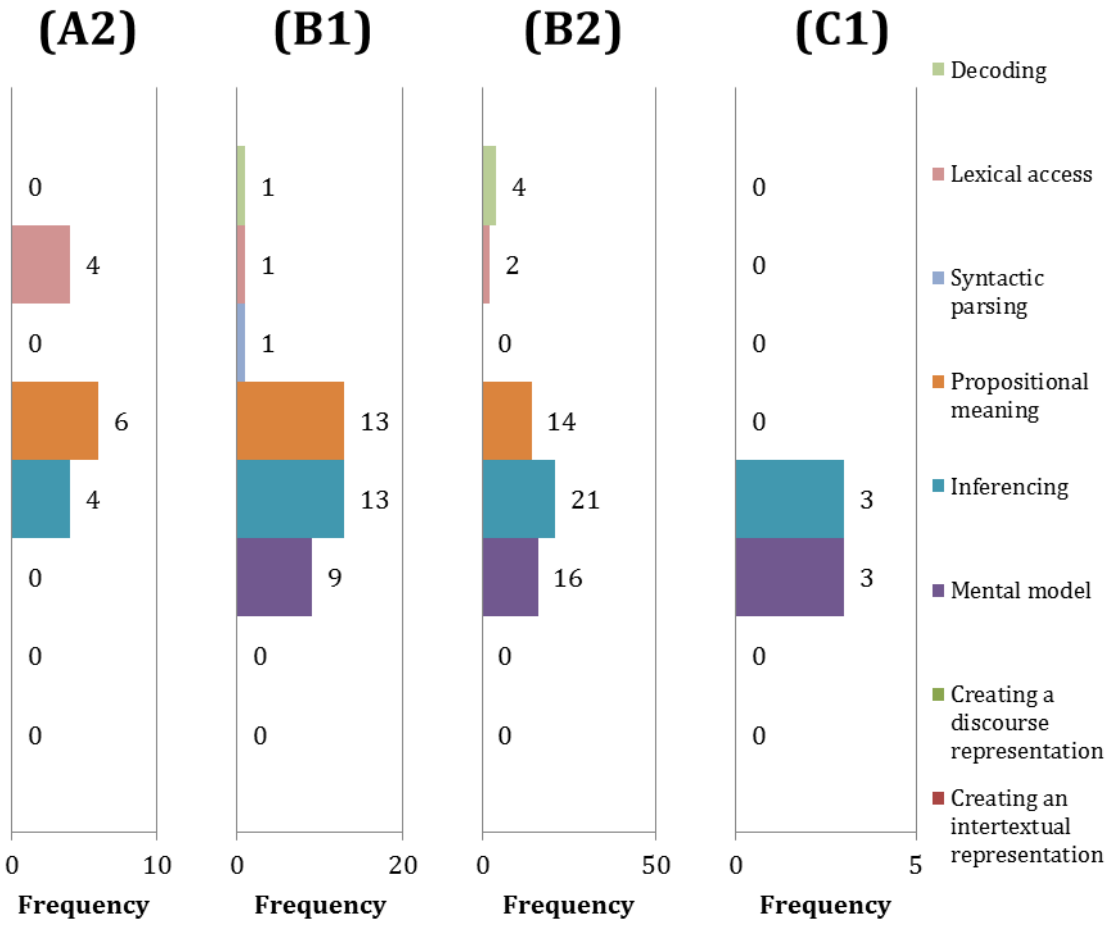


Figure 14 Highest level of cognitive processing (Listening)

5.1.3 Comparability of interpretations – to what extent the exam results are used for the same purposes

As described in European Commission/EACEA/Eurydice (2015), the majority of tests under consideration report results in terms of the Common Reference Levels of the CEFR. Igt should be the case then that users of the test results can interpret their meaning in terms of the the Can Do statements contained in the CEFR.

The large numbers of test providers reporting results in terms of the levels of the CEFR might lead us to expect that, in terms of inferences to be made, most exams targeted at the same CEFR level are directly comparable. The current report, however, cannot take successful targeting for granted. Jurisdictions may interpret results with reference to an external levels framework such as the CEFR but understandings of the CEFR levels may differ signficantly across jurisdictions. Considerable work is required to set exam standards (grade thresholds) and maintain them at a comparable level of difficulty across test forms and administrations (Council of Europe, 2001, 2009; North & Jones, 2009). In order to thoroughly appraise test alignment, extensive research would have to be done, involving active participation of all test providers. This, however, is beyond the scope of the current study. It was, however, possible to analyse some information relating to alignment to the CEFR, which is described below in section 5.2.

Among the few tests for which results are not reported in terms of the CEFR, the most important distinction in terms of the interpretation of results is whether they are norm or criterion referenced. Strict norm referencing would produce a result where at each test administration a fixec proportion of candidates are categorised within a oarticular grade (e.g. pass/fail). Any interpretation of what this may mean in practice is therefore relative to the other candidates in the same results group. Criterion referencing attempts to relate results to external criteria, such as the Can Do descriptors of the CEFR. These provide users of the results with a means to interpret the results in relation to real-world parameters, rather than the scores of a large group of candidates.

Figure 15 shows, where the information was provided, the extent to which tests which do not state an explicit alignment to the CEFR have results referenced to other criteria. This is important for any future attempts to compare results, as it may be possible to align different referencing systems based on criteria, and hence compare the results of the tests. However, where norm referencing is used, the prospect of comparison is more remote, as the characteristics of the each sample of candidates becomes significantly more important. As the charts show, where information was provided, criterion referencing was far more common than norm referencing, and, as such, suggests a slightly better chance of comparing the test results concerned.

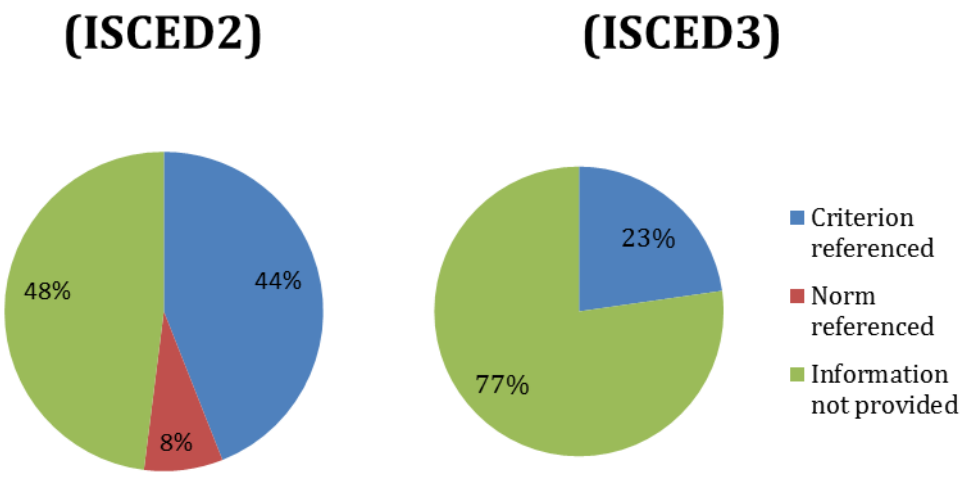


Figure 15 Interpretation of test results which which do not state an explicit alignment to the CEFR

5.1.4 Comparability of populations – similarity between candidates taking the exams

When comparing test results, assessing the comparability of the candidates taking the tests is important, especially because the target candidates under consideration in the current study will be expected to experience rapid cognitive development this stage in their lives. As Kolen and Brennan (2004:434) argue, 'two tests might measure essentially the same construct (at least in a general sense) but not be appropriate for the same populations'.

Table 9 summarises the age distributions at ISCED 2 and 3, according to the information provided. The median age suggests that candidates tend to take ISCED 2 exams at around 15 years old, and ISCED 3 around three years later. This difference suggests that the content of the exams is likely to be orientated differently towards each age group. In terms of content, this argues against any comparison across ISCED levels (however, see section 5, where we link the two levels to the CEFR, and thus to each other in terms of a notion of difficulty). Within levels, however, more comparability can be expected. The spread (as measured by the Standard Deviation (SD)) is around 1, such that most ISCED 2 candidates will be between 14 and 16 and most ISCED 3 candidates between 17 and 19.

Table 9 Candidate age distribution summarised

	mean	median	mode	SD	range
ISCED 2	14.69	15	14	1.24	6
ISCED 3	17.73	18	18	0.86	3

5.1.5 Measurement characteristics and comparability – facets which may make test results unreliable

There are many features of a test which potentially add to measurement error. The presence of such error means that the results of any test, their interpretation and, therefore, the comparability between tests suffers.

It is not straightforward to assign measurement error directly to specific causes. However, information about the personnel and procedures involved, and examination of sample materials can show likely sources of error. Such an investigation is within the scope of this research and is reported in the current section. Because, in most cases, the same test provider is responsible for tests at both ISCED levels, it was thought likely that there would be no meaningful distinction between ISCED levels when reviewing this data.

Personnel

The personnel involved in test provision are an important consideration, because, as well as being responsible for the correct conduct of procedures, they are needed to provide their expertise at crucial points in the process. For this reason, suitable selection, training and monitoring procedures must be in place.

Where information was provided, more of those involved in test construction were recruited according to specific criteria, rather than current employment status (e.g. currently a teacher in the education system) (Figure 16).

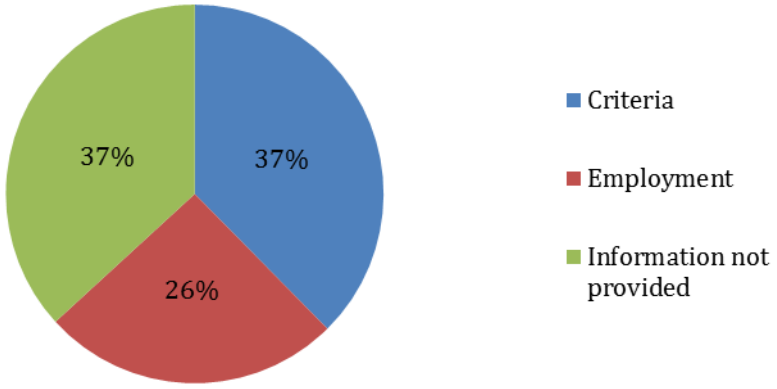


Figure 16 Reason for selection of test constructors

Figure 17 summarises the criteria stated as being used in recruitment of test constructors, where this information was available. A combination of criteria relating to education and work experience (either teaching or other work experience), or only work experience (as a teacher or tester) were the most common criteria. Use of quality standards specified by the Association of Language Testers in Europe (ALTE) was recorded in the case of one provider. This suggests a more through approach to recruitment but, given the scope of the current research and the response rate to requests for information, it is impossible to report more.

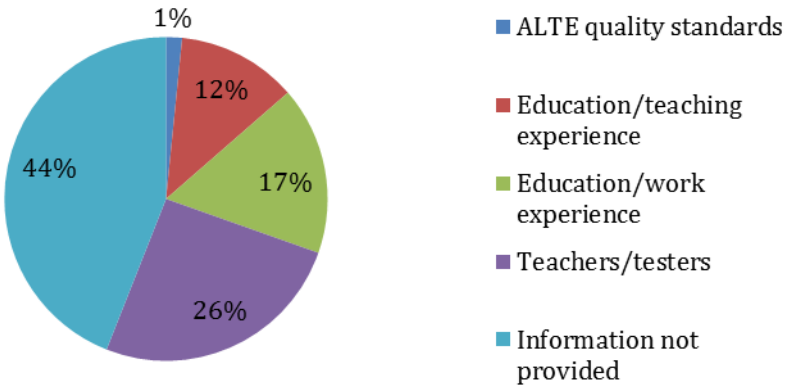


Figure 17 Selection requirements for test constructors

Like those involved in test construction, recruitment of those responsible for marking or rating candidate responses is also important. As Figure 18 shows, in most cases, information relating to the recruitment of markers and raters was not supplied. Where information about selection was provided, current employment status was more than twice as common as the use of specific criteria. This suggests the likelihood of greater variability between markers and raters, as it does not prioritise relevant expertise and suggests that there is no procedure to remove poorly performing staff.

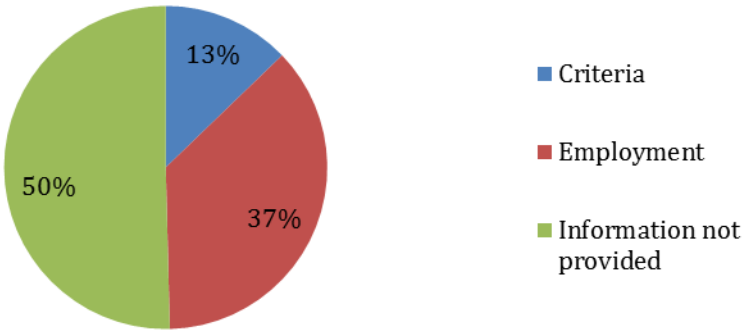


Figure 18 Reason for selection of markers and raters

Figure 19 shows that, where information was available, experience as a teacher or tester is far more important than a combination of education and experience, or the use of criteria defined as relating directly to quality (ALTE standards).

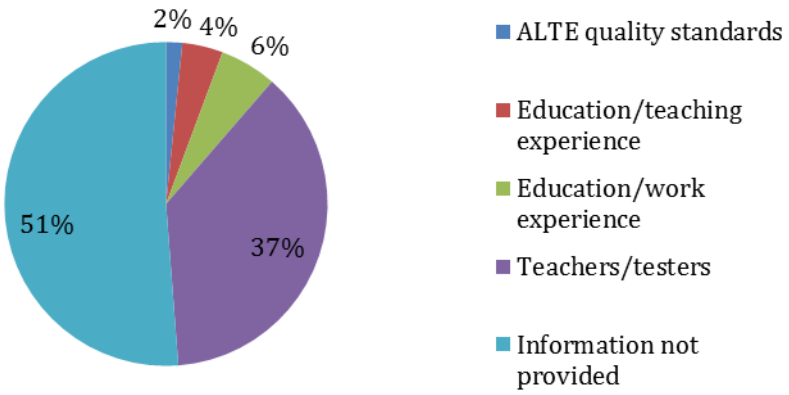


Figure 19 Selection requirements for markers and raters

After recruitment, a stage of training for markers and raters is considered important. As well as helping them to understand how to conduct the process, for raters in particular, this can be useful in ensuring they interpret the rating criteria as intended, rather than in non-standard ways (ALTE & Council of Europe, 2011). Standardisation usually provides raters practice in using the rating criteria so that they and their employers are sure that they can use it correctly. As Figure 20 and Figure 21 show, these questions drew a very poor response rate. As a consequence, it is hard to be sure whether lack of training and standardisation represent a serious threat to comparability.

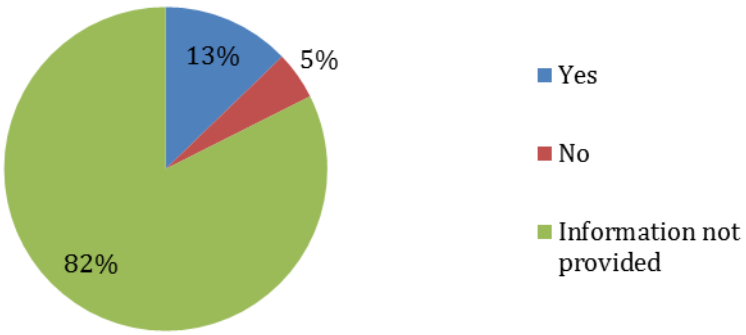


Figure 20 Provision of training for markers and raters

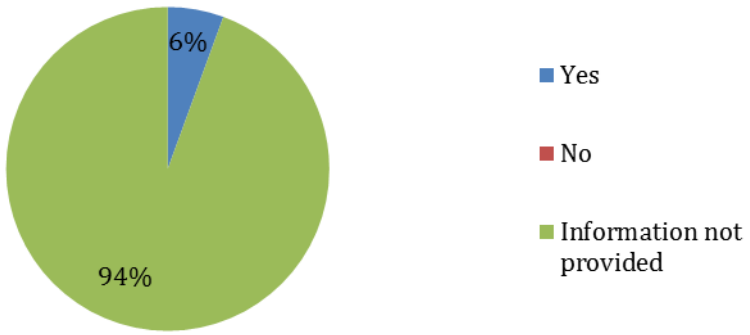


Figure 21 Provision of standardisation for markers and raters

Procedures

In order for results to be comparable across tests, they must first be comparable for all the forms or administrations of the same test. Standardised procedures help to ensure that individual tests yield comparable results no matter which form or administration each result comes from. Such procedures are of considerable importance, as it makes little sense to assess the comparability between tests which are not stable. Some procedures are aimed at reducing measurement error in specific parts of test provision, something which enhances stability and comparability. One important area where this can be done is in scoring or rating test performances. In all cases, as ISCED 3 tests typically have higher stakes attached to them (European Commission/EACEA/Eurydice, 2015), it was thought better to analyse the ISCED levels separately.

Procedures: Scoring performance tests

The process of scoring Speaking or Writing performances usually involves asking an expert to apply rating criteria of some kind to the performance. As human judgement is susceptible to various kinds of effects, such as severity or inconsistency, if only one rater is used, it is hard to be sure that such effects have been avoided.

Although there are a variety of approaches to detecting and/ or avoiding unwanted rating effects, perhaps the easiest and most common way to implement this is to ask more than one rater to score performances. Figure 22 shows that very little information was made available about the number of raters used to judge Speaking performances. Where information was available, one rater was most common, with two raters being relatively rare.

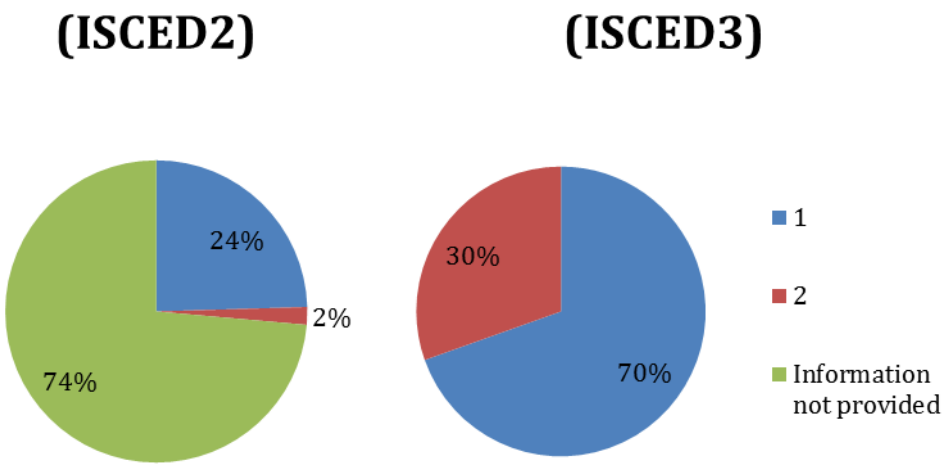


Figure 22 Number of raters per Speaking performance

Figure 23, which shows the number of Writing raters used, shows a similar picture to that of Speaking. Slightly more is known and the proportion of tests judged by two raters is higher but the number of exam performances judged by a single rater is high among those tests for which information was received. In the case of both Speaking and Writing, there is reason to believe that rating processes could yield a significant amount of error.

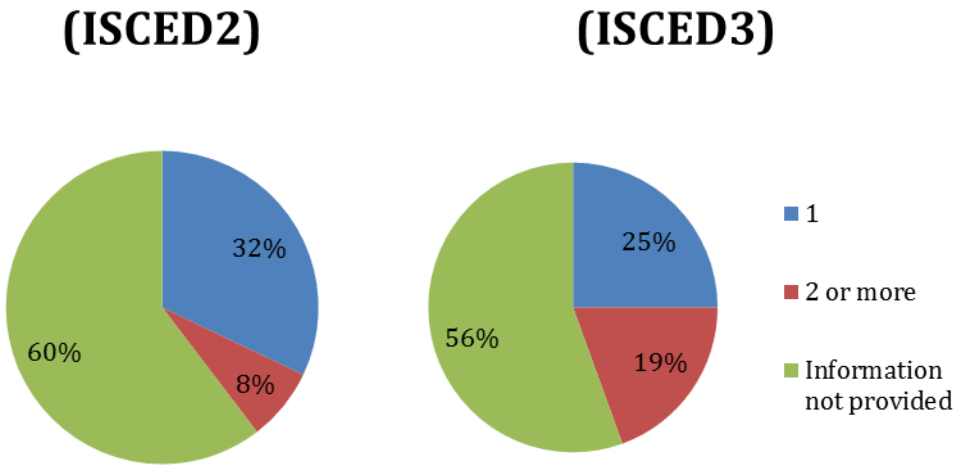


Figure 23 Number of raters per Writing performance

If test providers used more than two raters, they were asked whether there was an established procedure to resolve disagreement. However, because so few responded to requests for information about the number of raters, the information was too sparse to be of interest.

Procedures: Standards setting and maintenance

A crucial part of ensuring comparability between forms and administrations of the same test is the setting and maintenance of standards, or boundaries between grades (such as A2, B1 and B2 or pass and fail). Boundaries must be placed so that any guidance on how to interpret each grade accurately describes the ability of candidates within that grade. Furthermore, the position of the boundary must be maintained, so that guidance applies no matter which test form or administration applies.

The task of ensuring that grade boundaries fall where intended is made considerably easier if the test form used is the same or as similar as possible for each administration. As a result, test construction has an important role to play in obtaining equivalent results. Figure 24 shows responses to the question of whether an attempt was made to use the same or very similar tests for the same exam session. As the charts show, more than 75% of tests administered with this provision, which is likely to aid comparability between tests.

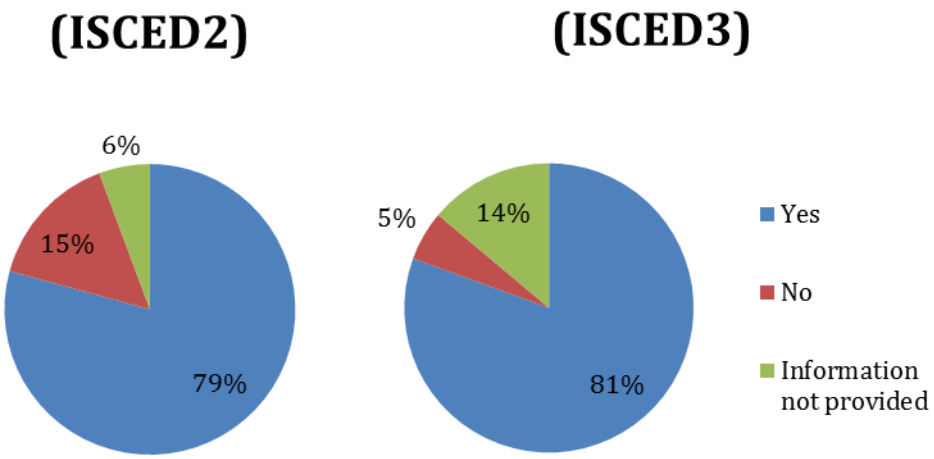


Figure 24 Provision of equivalent tests for all candidates in a single session

When asked if an attempt was made to ensure standardised tests for all candidates, even those not in the same session, less information was available. Figure 25 shows that information was not available for more than half of the tests in the study. Among those tests where information was available, more than half attempted to standardise across test sessions. However, a significant group did not and the proportion of information which was not available indicates that relatively little is known about whether the results of these tests can be compared at all.

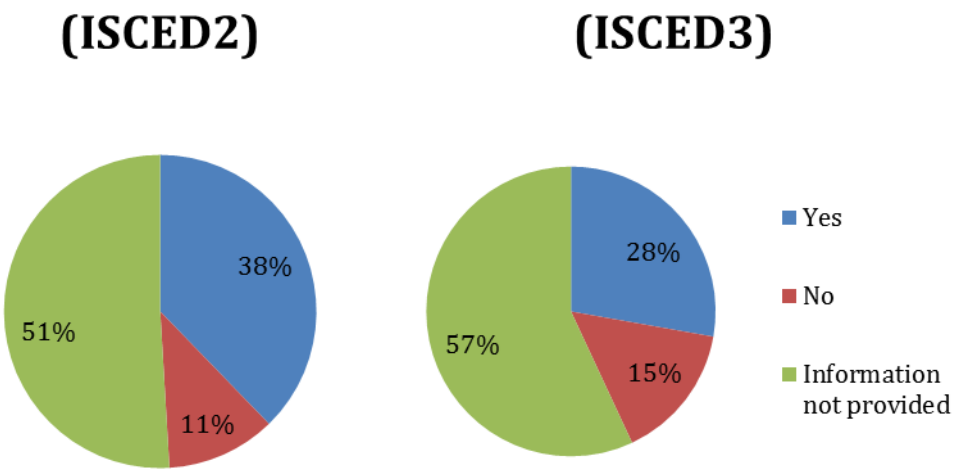


Figure 25 Provision for equivalent tests for all candidates regardless of session

The report "National Language Tests" (European Commission/EACEA/Eurydice, 2015) found that most tests under consideration are reported in terms of the CEFR. As was mentioned in section 5.1.3, however, it is possible for a test provider to state that the results of a test can distinguish between different CEFR levels even though there is little evidence to show that it does. Such statements should not be taken at face value given the differences in how the CEFR levels are understood . Within the scope of the current study only a limited validation of test providers' judgements could be provided. Test providers were asked how the alignment between tests and the CEFR was made. In many cases, as Figure 26 shows, this information was not made available. However, of those who did respond, most used Can Do statements without the addition of research procedures of the sort contained in the Manual for Relating Language Examinations to the CEFR (Council of Europe, 2009). This suggests that, in many cases, the basis for alignment is weak, and therefore comparability is not as evident as is implied by the findings of European Commission/EACEA/Eurydice (2015).

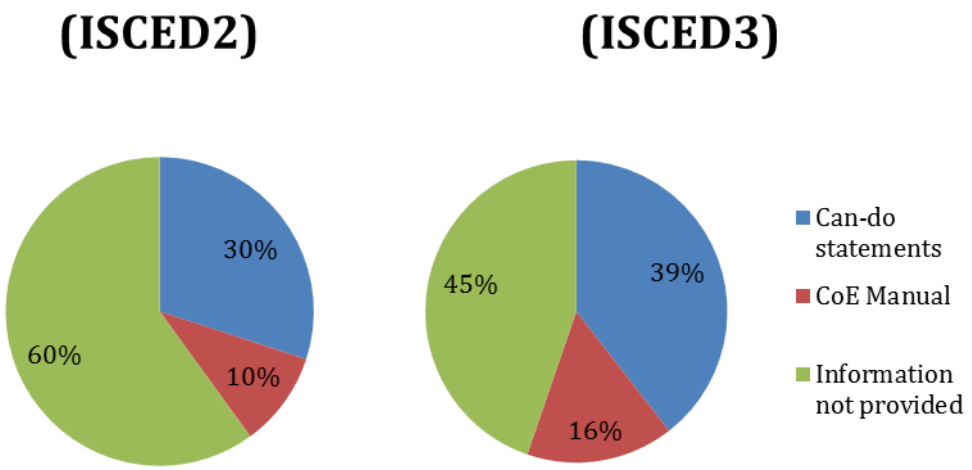


Figure 26 Approach to aligning the test to the CEFR

In addition to examining the way in which tests were aligned to the CEFR, where sample test materials were available, experts were asked to assess the level for each component. This was then compared to the level(s) the test was stated to report on. Based on a single test form in each case, Table 10 shows how well tasks in the test materials targeted the levels intended. For each skill at each level, a percentage is given for:

- the percentage of tests where the level of some tasks did not match the indicated level(s) of the test
- the percentage of tests where the indicated level(s) of the test was not found in any tasks
- the percentage of tests where the indicated level(s) of the test was found in at least some tasks.

Table 10 Targeting of tasks on CEFR levels

		A1	A2	B1	B2	C1	T
Speaking	% where actual level of test not as stated	0.00%	10.00%	6.25%	8.33%	0.00%	7.32%
	% where stated level not represented by tasks	100%	20.00%	37.50%	16.67%	100%	31.71%
	% where stated level represented by tasks	0.00%	70.00%	56.25%	75.00%	0.00%	60.98%
Writing	% where actual level of test not as stated	0.00%	20.00%	23.53%	3.70%	0.00%	15.48%
	% where stated level not represented by tasks	100%	25.00%	38.24%	59.26%	100%	44.05%

	% where stated level represented by tasks	0.00%	55.00%	38.24%	37.04%	0.00%	40.48%
Reading	% where actual level of test not as stated	0.00%	0.00%	19.44%	8.33%	58.33%	15.74%
	% where stated level not represented by tasks	100%	40.91%	30.56%	0.00%	8.33%	21.30%
	% where stated level represented by tasks	0.00%	59.09%	50.00%	91.67%	33.33%	62.96%
Listening	% where actual level of test not as stated	0.00%	0.00%	14.81%	13.64%	66.67%	14.29%
	% where stated level not represented by tasks	100%	25.00%	14.81%	22.73%	0.00%	20.78%
	% where stated level not represented by tasks	0.00%	75.00%	70.37%	63.64%	33.33%	64.94%

As can be seen in Table 10, there were considerable issues in targeting tasks as intended. As a result, the way in which results are interpreted, whether in attempts to compare them across tests, or for the kinds of uses the tests were designed for, are highly problematic. These findings suggest that the tests examined in the current study are not likely to be comparable with each other.

Reliability, denoting the accuracy and consistency of Speaking or Writing raters, or of test scores, is an important indicator of the quality of a test system. Poor reliability implies poor comparability. Estimating reliability is therefore a useful first step in managing test quality (University of Cambridge ESOL Examinations, 2011).

Figure 27 shows that very little information was available about the estimation of reliability. Very few test providers were found to estimate rater reliability, although slightly more did this for test scores. When the findings concerning rater reliability are considered together with the findings on the rating process, it must be concluded that there is little evidence to suggest that these processes are well controlled and yield accurate and reliable results.

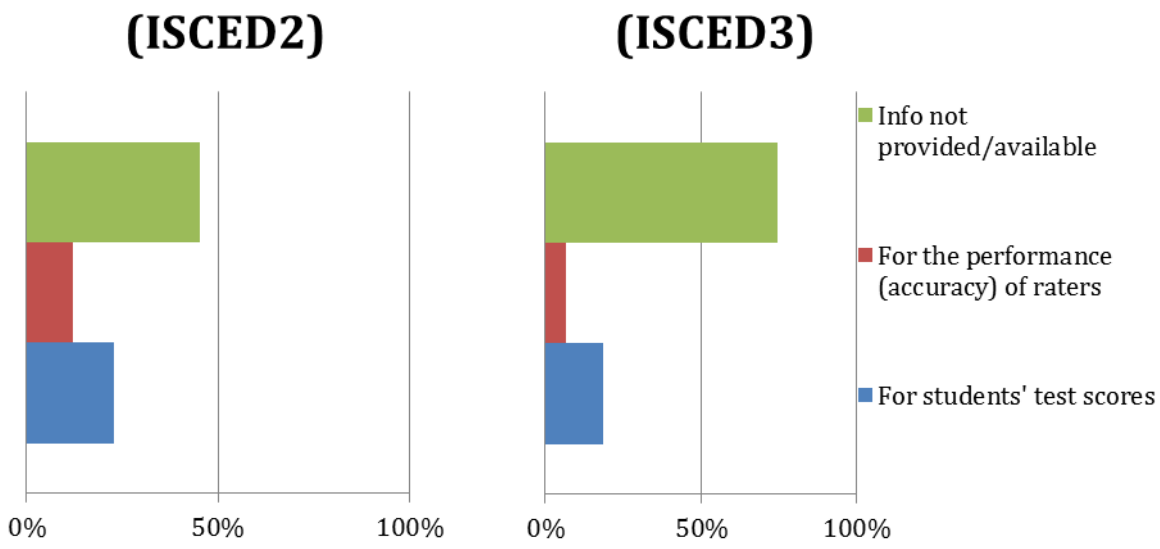


Figure 27 Estimation of reliability

Procedures: Materials

Test materials are the result of a production process which, in the simplest case, involves an individual writing all test materials for each form of the test. Review, pretesting, editing and item banking stages can all help to improve the quality of test materials. Within the scope of the current study, looking in detail at the test production processes adopted by test providers was not possible. Instead materials were reviewed, as these may provide indications as to where the process is inadequate. Flaws found in materials add measurement error to test results and make comparability between tests more difficult to achieve.

Test materials also control the way in which the test construct is measured. For example, these include different item types and differences in the way in which Speaking or Writing tasks are framed. Differences here are not flaws but represent a challenge to comparability because these contextual features mean that measurement of the construct is not quite the same across tests.

Procedures: Flaws in test materials

Test materials for Reading and Listening components are predominately made up of items which require short responses from the candidate, or the selection of a response from a range of options. For these kinds of items, the use of expert judgement in their design is exceptionally important. Poorly designed items yield little or no useful information about candidates and do not contribute anything meaningful to test results. Instead, they add to measurement error. For example, an item which requires world or cultural knowledge to be answered, even though the test is a language test, will discriminate between candidates on grounds which are irrelevant to the aims of the test. Such errors in item writing are likely to be randomly distributed among items, rather than systematic, so this source of error is also a threat to comparability.

Figure 28 and Figure 29 show the range of errors found amongst the Reading and Listening test materials examined. In both cases, a considerable number of flaws were

identified by expert analysts. Without response data, it is impossible to tell to what extent they affect the performance of candidates on items and thereby contributes to error. In can be concluded that item flaws are likely to form a considerable proportion of measurement error, and therefore reduce comparability.

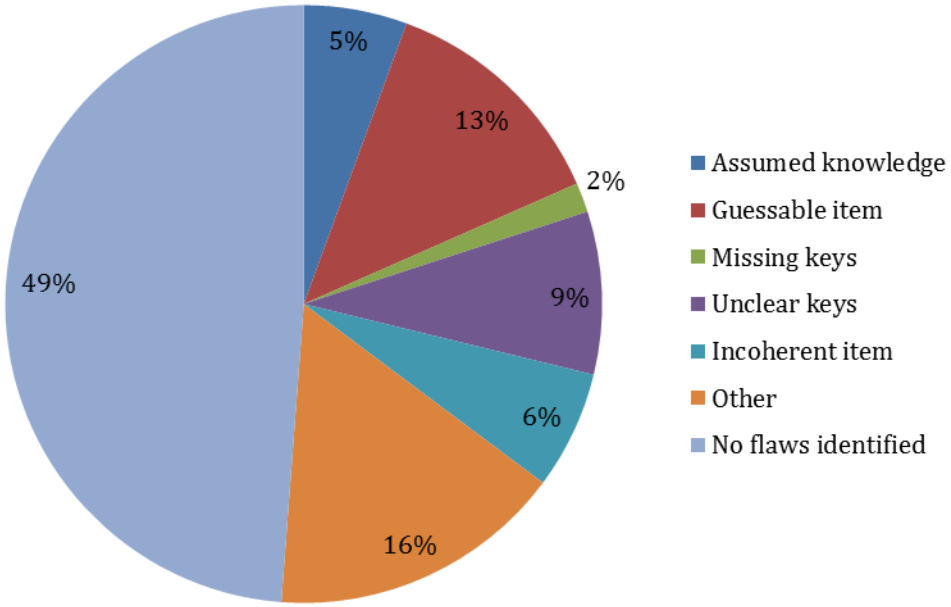


Figure 28 Flaws apparent in sample Reading materials

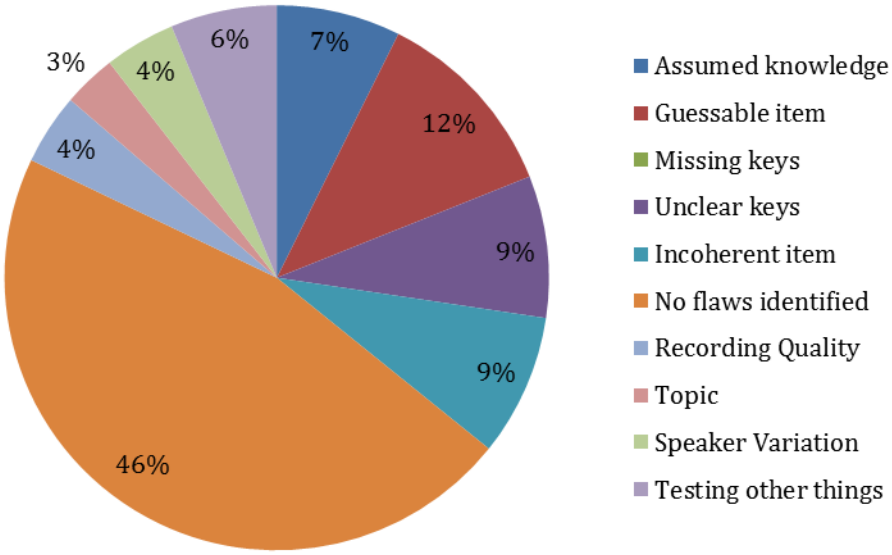


Figure 29 Flaws apparent in sample Listening materials

5.1.6 Equating English and French standards

We considered two ways of establishing the relative levels of English and French:

1. Using the ESLC data which defined task difficulties and CEFR level thresholds as described above. The standard-setting conference set standards on a by-language basis, although there were also post-hoc checks aimed at verifying the comparability of standards across languages.
2. Using an explicit link in the current study between the English and French tasks, which is provided by two dual-language datasets (for Reading and Writing) in which both languages were included, enabling raters with competence in both languages to rank these tasks relative to each other.

The second approach was chosen as being more direct, and making best use of the options which CJ offers us. English was taken as the point of reference and French was linked to it using a chained procedure via the dual-language datasets: English ➔ dual-language ➔ French.

This involves scaling the dual-language data to the English using a mean-and-sigma method, and then applying the updated dual-language scale to align the French scale with the English.

The charts below show the relationship of English to the dual-language anchor, and of the anchor to French. The points in these graphs are the test tasks calibrated within the No More Marking Comparative Judgement website.

As explained further, it is important to use these plots as a visual check on the nature of the alignment. There is some evidence of stretching in the lower tail of the CJ data,

due to some tasks being found as very easy. This informed a decision to remove 3 or 4 of the easiest data points from the anchor.

It is worth pointing out that the dataset which constitutes the anchor forms a different set of links with the English and French datasets – the larger number of points in the link from English to the anchor is because there are more English tasks in the anchor than French ones.

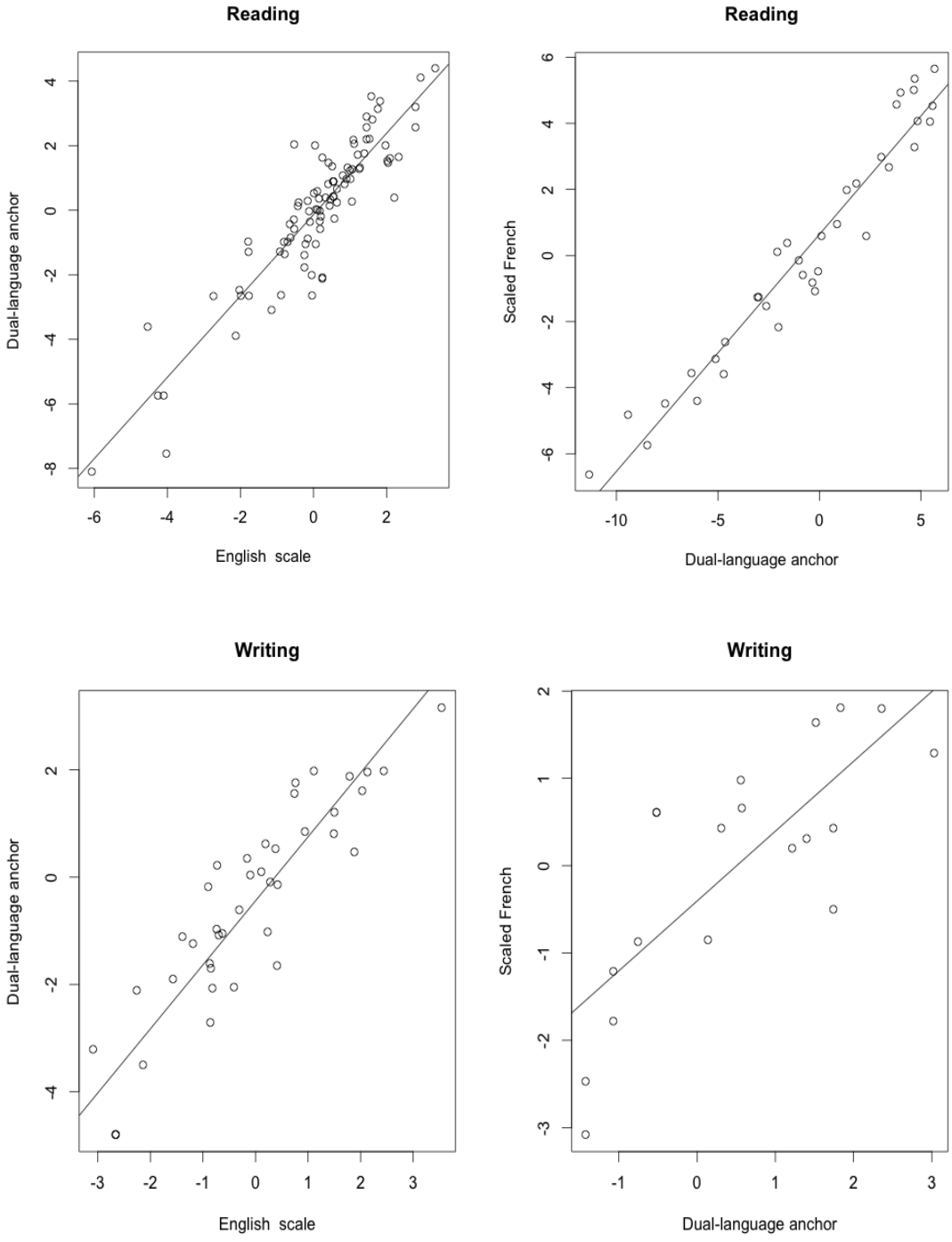


Figure 30 Linking English to French via a dual-language anchor

Effects of the test context

Item types were examined for components which had an explicit lexico-grammatical focus, so-called linguistic competence tasks. Such tasks are difficult to place on the scales of the CEFR, as they do not target one of the four skills and, as such, there are limited applicable Can Do descriptors. These components are, none the less, still worthy of consideration and are quite likely to include the largest variety of items types among all components. Figure 31 shows that, at ISCED 2 level, a smaller variety of items types are employed than at the ISCED 3 level. As a result, linguistic competence components in ISCED 2 tests are likely to be more comparable with each other. The reasons for this distinction are likely to be due to the greater prevalence of this component at higher levels of competence (likely to be found at ISCED 3) and the need to find item types which concentrate on more specific testing foci.

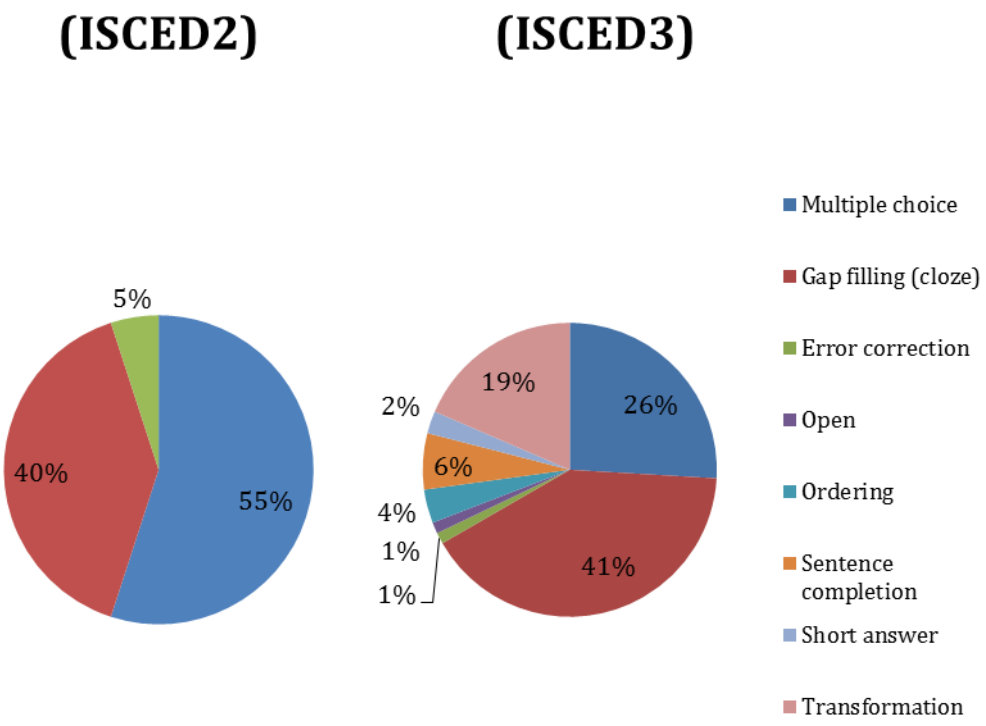


Figure 31 Item types used in tests of linguistic competence

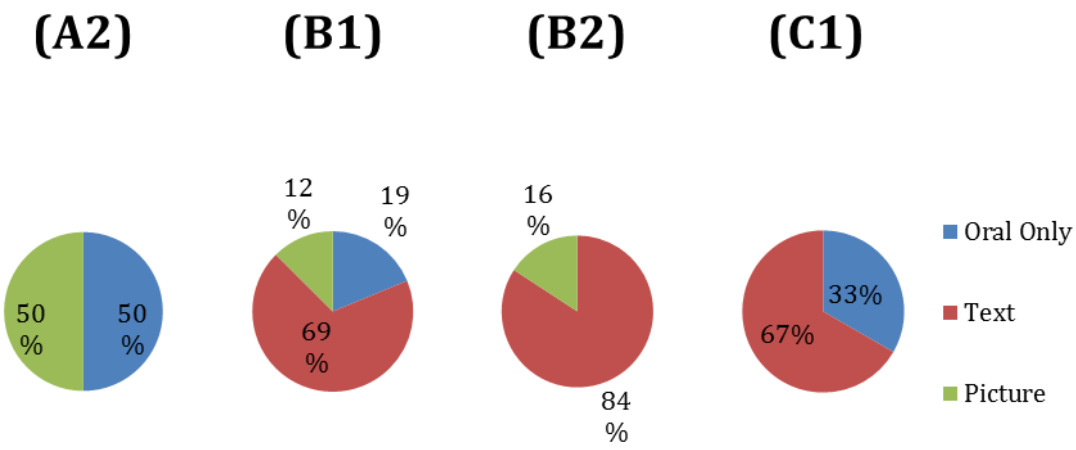


Figure 32 Mode of task prompt (Speaking)

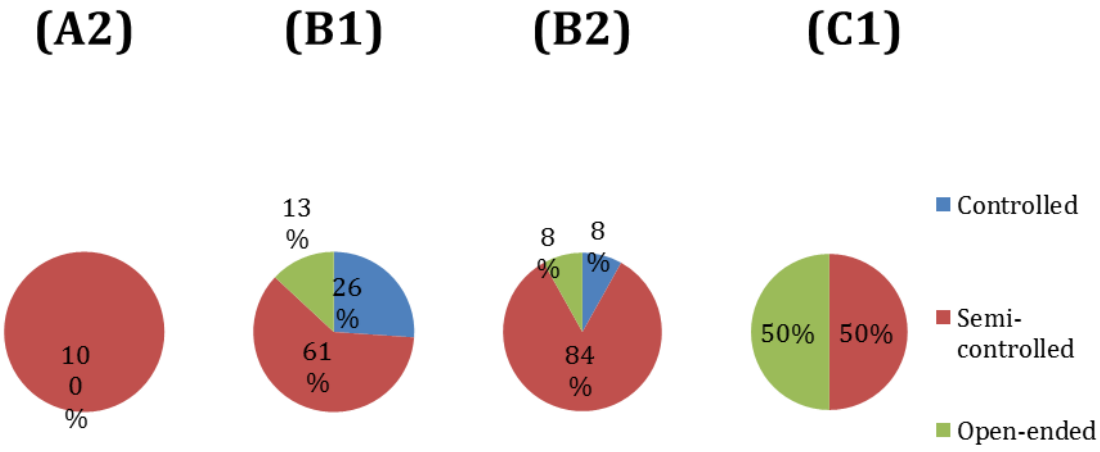


Figure 33 Control/guidance (Writing)

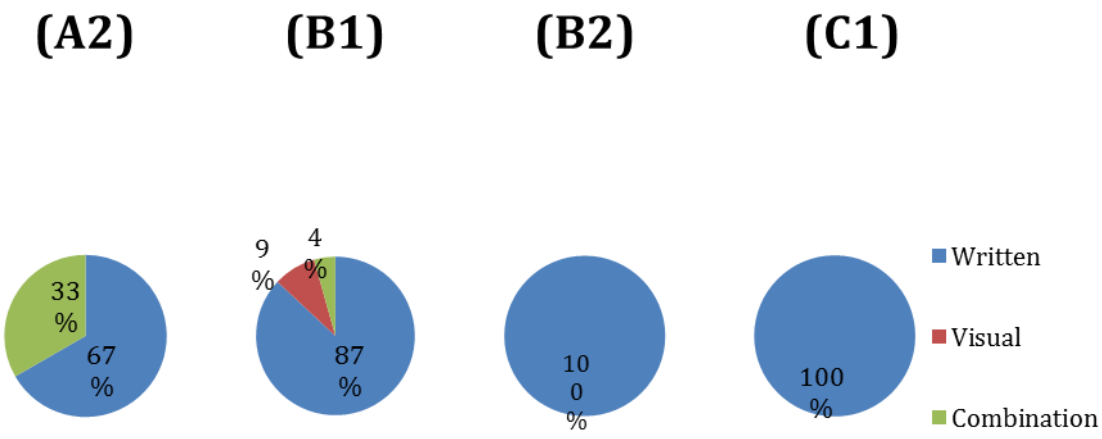


Figure 34 Mode of input (Writing)

In Figure 32, Figure 33 and Figure 34, a selection of features of the task setting for Writing and Speaking are presented. These affect the way in which tasks are controlled and, although not as important as comparability between constructs or the setting and maintenance of standards, differences affect the comparability between tests.

The first refers (Figure 32) to the way in which a prompt is provided to candidates to elicit speech. The second (Figure 33) is the level of control implied by the instructions and/or information provided in a Writing task. Candidates can be given content to be included in their response, told the number of words to write, or told who to address, for example. At one extreme, they may be given hardly any guidance, at the other, it may be strict. Control can help to support weaker candidates in constructing their responses but also provides more comparable samples which then leads scores which can be more easily compared from candidate to candidate. In all cases, differences between tests are likely to elicit quite different performances. Finally, the way in which the stimulus for Writing tasks is presented is summarised (Figure 34).

As might be expected, the use of pictorial stimuli is more common at lower ability levels. Differences do not, however, seem entirely due to level, as the use of text and oral prompts in Figure 32 does not form a pattern across the levels. In Figure 34, written input predominates at all levels. As a consequence, there are clearly important differences between tests which will hamper comparability. In Figure 33, which deals with control or guidance in Writing, as might be expected, less control is deemed to be required at higher levels. This is probably a result of the reduction of 'scaffolding' to support the performances of weaker candidates. These results, none the less, suggest another source of variation which affect the comparability of performances, and therefore results across tests.

5.2 Quantitative analysis: the Comparative Judgement exercise

The Comparative Judgement exercise presents an approach through which tests may be compared, based on comparative judgements of the samples of students' performance for Writing and Speaking. As explained in section 2.3 above, samples of performance were not available for this study. However, in order to demonstrate the potential of this technique for future comparability studies, Reading and Writing tasks were included in this exercise and ranked on the basis of their difficulty. By seeding tasks where the CEFR level is already known, the resulting scale may be anchored to the CEFR. However, in order to fully assess comparability a further stage is needed, requiring information on distributions of test scores for a given exam/jurisdiction.

		English		French	
	All	Reading	Writing	Reading	Writing
Number of tasks judged	204	101	40	44	19

Figure 35 Numbers in the Comparative Judgement exercise

Number of judges taking part: 49

Number of decisions per judge:

- minimum 1
- maximum 4,800
- mean 705
- median 50
- mode 50

The rater who completed 4,800 judgements was unusual, as the median and mean scores indicate. But this large total does not necessarily indicate a threat to quality: the intention of Comparative Judgement is that large numbers of judgements can be made in a short space of time.

5.2.1 The link to the first European Survey on Language Competences

The first European Survey on Language Competences published its findings in 2012. The Survey was implemented on behalf of the Commission by the consortium SurveyLang led by Cambridge English Language Assessment Language Assessment. It tested the skills of Reading, Listening and Writing in 16 participating countries or jurisdictions, for first and second foreign language, from the tested languages English, French, German, Italian and Spanish. Speaking was considered too logistically difficult to include.

Despite this partial participation, the survey provided a valuable first insight into standards achieved in the participating countries and jurisdictions. What was striking was the wide range of achievement observed, and its frequently quite low level, given the years of study involved. This first orientation as to the standards being achieved may, we believe, be taken as a reasonably reliable point of reference for measuring future developments.

Thus the current study addressed carefully the issue of how to carry the standard forward. An approach was required to enable judgements made in 2015 to be anchored to those made at the standard setting conference for the European Survey which took place in Cambridge in September 2011, with the participation of language professionals from across Europe. A major goal of the comparative study was to construct a measurement scale against which all jurisdictions could be compared. The objects to be compared were the test tasks collected from participating jurisdictions, to which were added selected tasks from the European Survey, to act as anchors.

The comparative study required judges to decide which of successive pairs of tasks was the harder one. Tasks contain several items, but rather than compare tasks item by item judges were required to compare them according to an impression of overall difficulty.

Figure 36 is taken from the ESLC Final Report and illustrates the difficulty of the tasks used in the English Reading test. The tasks are ranged along a theta scale from easiest to hardest. The end-points of the line for each task represent the difficulty of achieving a 50% and an 80% score respectively, which are taken as arbitrary indicators of 'basic' and 'full' mastery of the task.

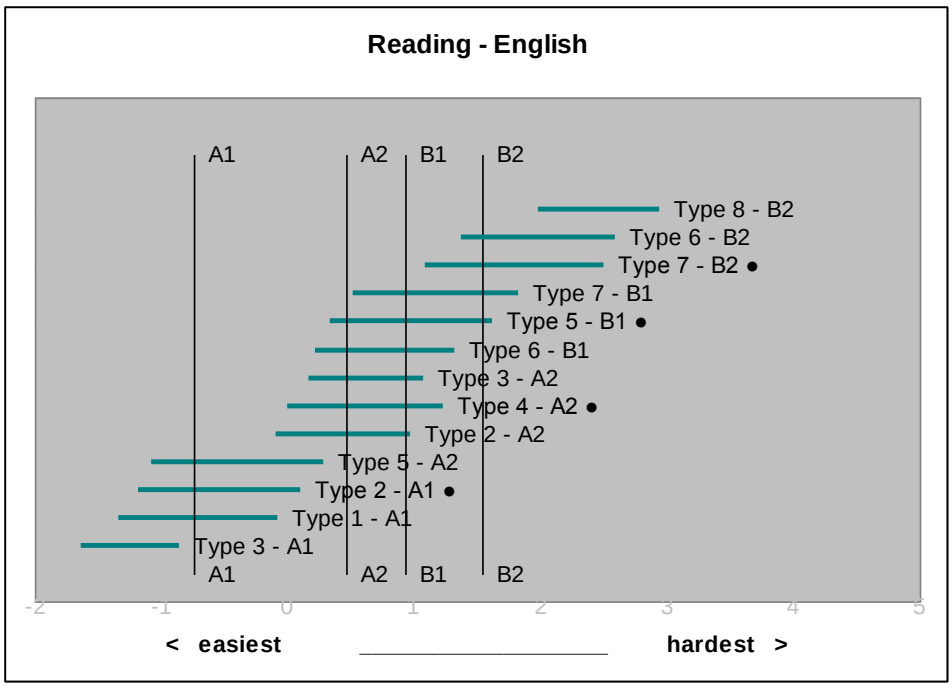


Figure 36 Example ESLC report showing task difficulty relative to CEFR levels

These charts were originally developed for use at the standard-setting conference, and were used in conjunction with the texts of the tasks themselves to provide a framework within which judges could determine CEFR levels. The levels resulting from these determinations were added to the charts for inclusion in the report.

From the point of view of this study 50% is a critical point of reference, because a score of 50% can be taken as the mean score against which the anchor tasks can be compared with the countries' tasks in the comparative study; it also relates usefully to

the notion of a threshold between a lower and a higher level - that is, the ability at which the chances of being assigned to the lower or higher level are 50-50.

These charts thus provide an approach to selecting a range of ESLC tasks capable of anchoring the scales emerging from the Comparative Judgement exercise to the scales reported by the ESLC. By anchoring the measurement scale in this way it should be possible to compare the two studies transparently.

5.2.2 Making the link to the European Survey on Language Competences

Section 5.2.1 above has already presented the forms of data available to link this study to the ESLC.

Practically this involves taking tasks from the ESLC and using them as anchors, taking the difference between their ESLC scale values and their values from the CJ exercise to perform a linear scaling, or rather, a number of linear scalings by group.

There are thus two stages in the approach to scaling adopted: first scale French to English via the dual-language anchors, and secondly link the scales to the ESLC via the ESLC anchor items.

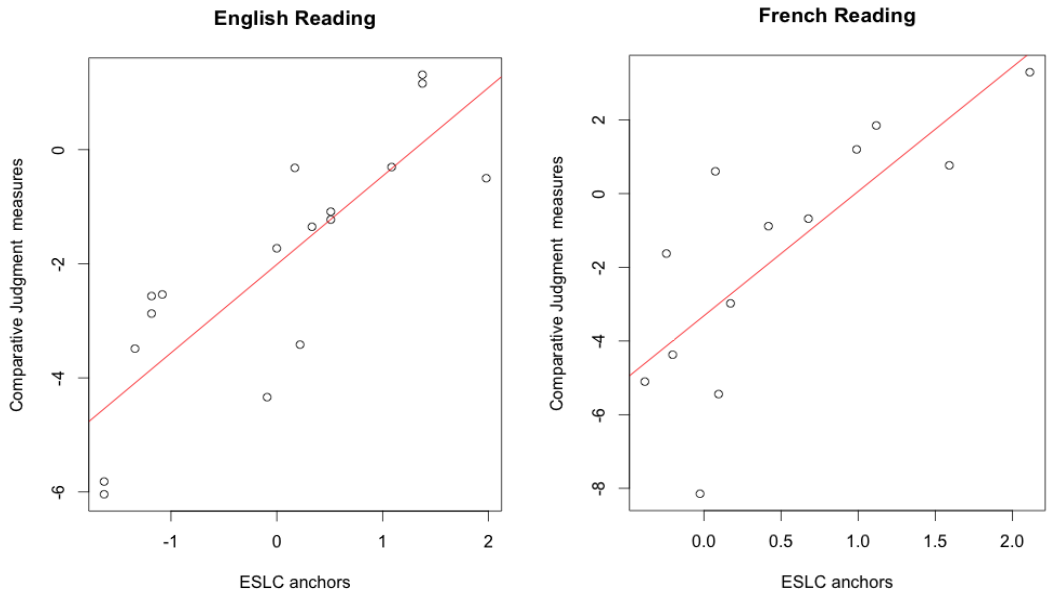


Figure 37 Comparison of the ESLC anchor tasks as a basis for scaling

Note that the correlation of the CJ outcomes and the original task calibrations is moderate. Note also that here, as in the case of the chained English-French linking reported above, there is evidence of stretching in the lower tail of the CJ scores. This prompted a decision to delete the 3 or 4 easiest items from the anchor.

5.2.3 Findings

The difficulty of tests at ISCED 2 and ISCED 3 levels

Figure 38 below shows the distribution of task difficulties estimated from the Comparative Judgement response data. The picture for English is clearer because of the larger number of tasks included in these datasets.

The distribution is shown against the CEFR levels as estimated from an anchoring to tasks from the ESLC. The upper figure is ISCED 2, the lower figure ISCED 3. The separation of the two distributions can be seen clearly for English.

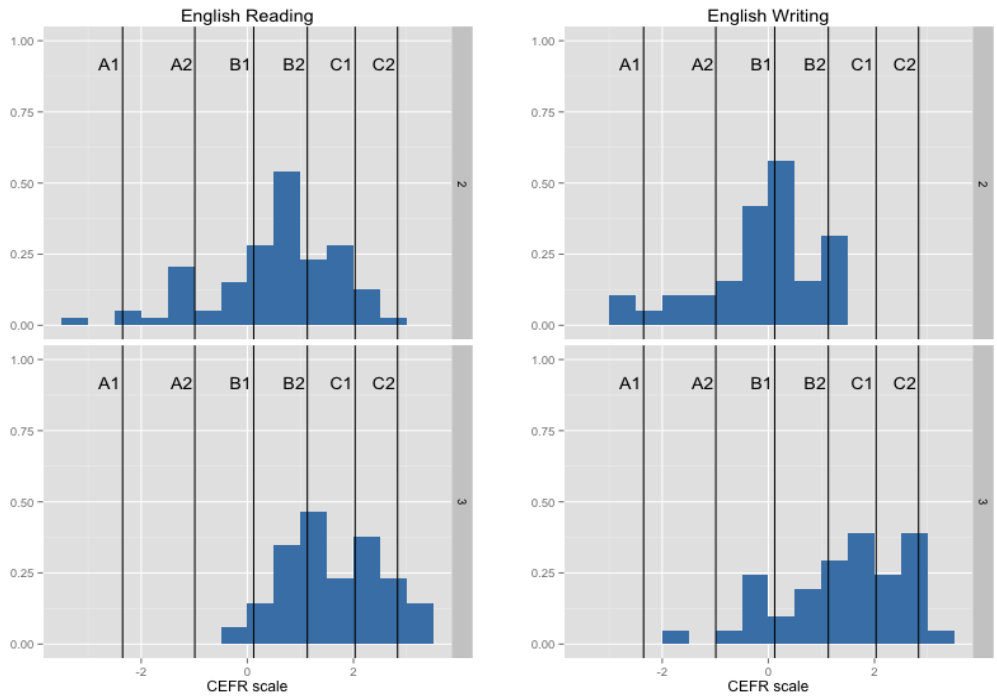


Figure 38 English Reading and Writing task difficulty

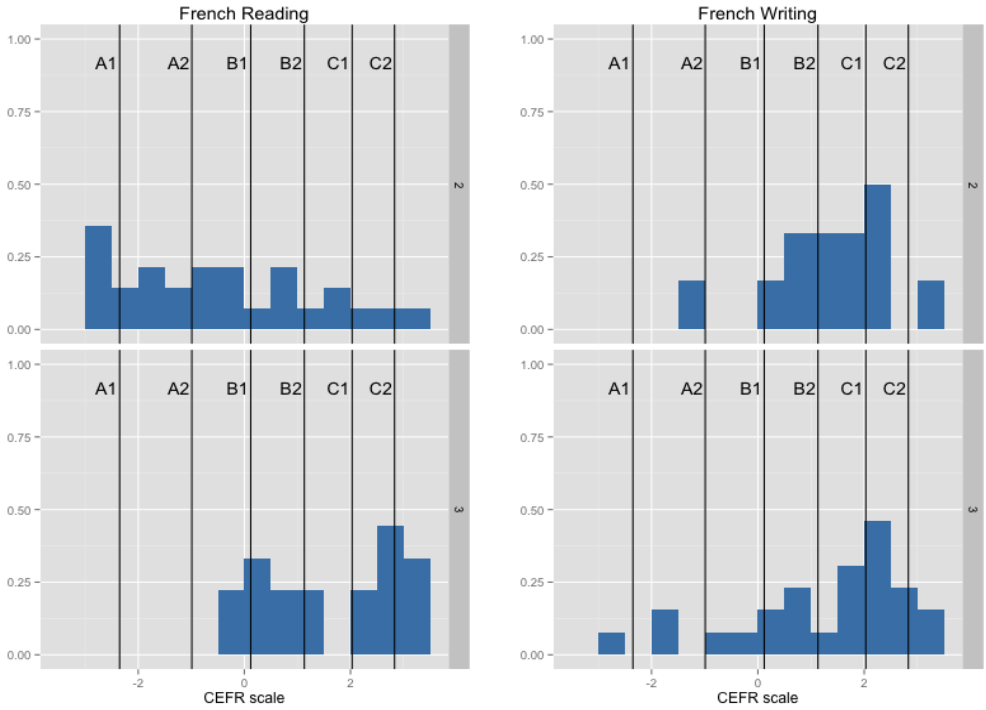


Figure 39 French Reading and Writing task difficulty

For French the ISCED levels are reasonably distinct for Reading, less so for Writing

An informal impression of tasks related to CEFR levels

An important goal of this project is to communicate the idea that a range of different jurisdictions' test tasks can be brought onto a common scale and that this scale can be given a real-world interpretation by referring it to the CEFR. The methodology demonstrated in this section 5.2, and illustrated with examples of test tasks for Reading and Writing, provides a strong psychometric basis of how national results of students' language competences could be mapped to the common scale of the CEFR.

Please treat the tables above as an informal communicative device, attempting to show how test tasks might function in CEFR terms. It is not a judgement about jurisdictions' achievement, because we cannot relate it concretely to student performances, and we do not necessarily know how each test was graded. A hard test with a low pass mark is the same as an easy test with a high pass mark. Moreover, in these data each jurisdiction may be represented by as little as one sampled test task: we cannot know the overall difficulty of the sampled tests.

While we can talk of a student being "at a level", depending which side of a threshold they stand, we cannot think of tasks in the same way: tasks just provide evidence about the student. To do this effectively they must be targeted at an appropriate level - for example, we write a B1 task so that a student at B1 level has, say, a 70% chance of answering it correctly. So the relation of a test task to a CEFR level is indirect and factors in how challenging we choose it to be (not too hard or too easy). So if a task appears in a table as a B1 task, this means "this is a good task to use at B1 level". These tasks should therefore be seen as useful indicators of the range of levels being covered by tasks developed for testing at ISCED2 and ISCED 3 level in national language tests in Europe.

5.3 Qualitative and quantitative: combining both studies

There are thus two main sources of evidence in this study. The qualitative evidence includes that provided by the Eurydice Network and additionally by Cambridge English Language Assessment, who engaged subject experts to further evaluate available documentation and in particular to undertake a rigorously-structured analysis of the test material. The quantitative evidence is provided by the Comparative Judgement exercise, which has enabled us to construct measurement scales and relate them to the CEFR, and to the European Survey on Language Competences.

In adopting both qualitative and quantitative approaches we may arrive at a richer understanding of the phenomena of interest. The combined evidence from these two sources should, we may hope, be more than the simple sum of the parts.

The amount of data produced by the two studies is considerable, and what is attempted below is an initial exploration of a sample of questions addressed within the qualitative survey. It produces some interesting outcomes.

The comparison with the CJ exercise is based on matching the qualitative and quantitative datasets by country and ISCED level. Easiest to compare are the categories used in describing the test tasks, particularly categories which are implicitly or explicitly ordered. Judgements relating to the CEFR levels are explicitly ordered. Judgements relating to traits such as level of cognitive challenge also represent explicit hypothesised progressions, but being presented in descriptive form may be more prone to different interpretations.

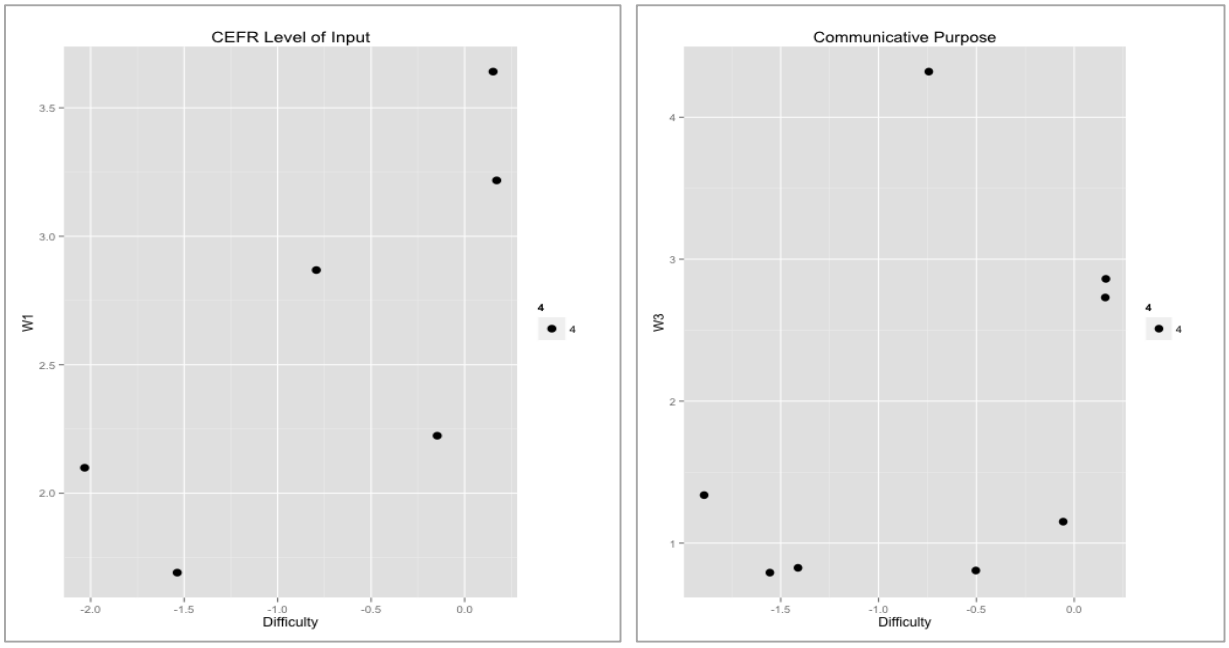


Figure 40 Two questions about English Writing at ISCED 2

Figure 40 shows the result for two questions concerning the skill of Writing. The hypothesis to test is whether jurisdictions whose test tasks have been found more difficult in the CJ exercise (on the x-axis above) are also judged more difficult by the expert analysts in the qualitative study (the y-axis). On the left is a question on the CEFR level of input of a Writing task. The expected positive (though weak) relationship is found, as the points indicate a diagonal from bottom left to top right.

The first obvious issue is that the number of data points for making these comparisons is quite small, given that the data are divided by ISCED level and language tested, and that many cells in the original spreadsheet are empty.

The right-hand plot shows a question about the communicative purpose of the Writing task. These are categorical options, ordered as follows:

- referential (telling)
- emotive (reacting)
- conative (argumentation, persuasion)
- phatic (social interaction).

It is not clear whether these options are intended to form an implicit progression, and the plot certainly does not suggest one - the points on the x and y axes do not correlate at all. .

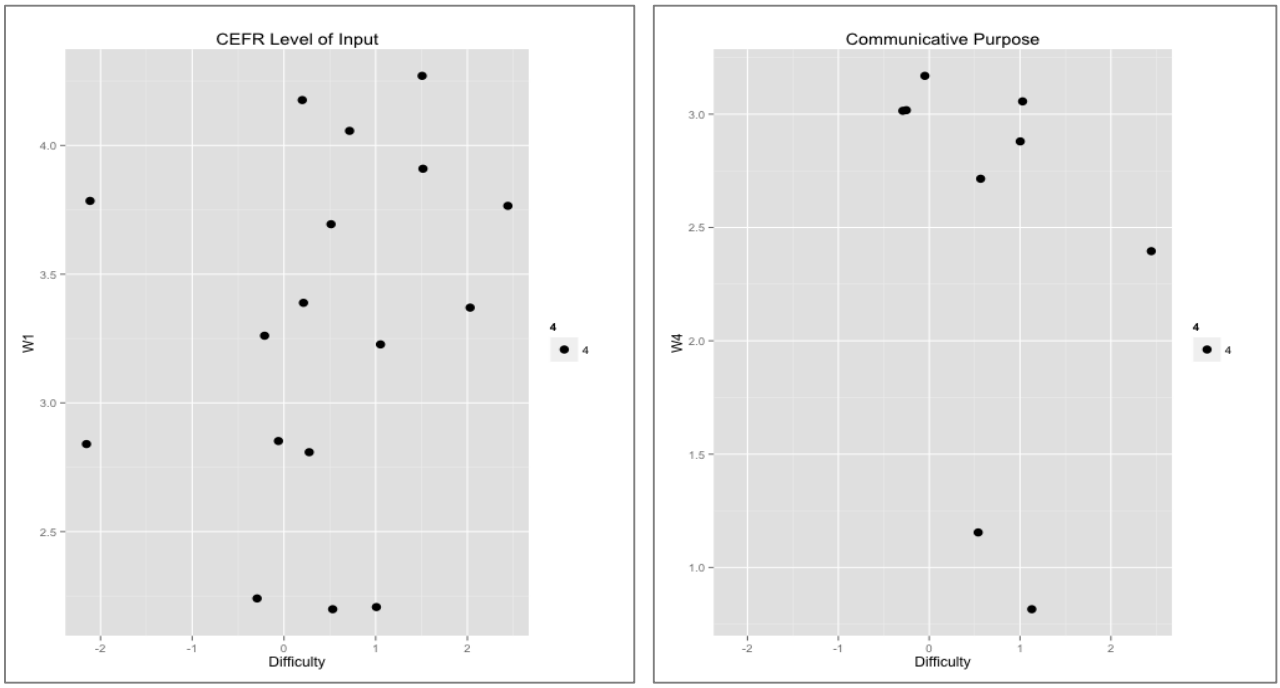


Figure 41 Two questions about English Writing at ISCED 3

In the figure above the left-hand plot shows that level of input does not appear to relate strongly to the difficulty of the ISCED3 task. This may be explicable, if the input text is not the primary feature of the task which determines its difficulty (which seems likely at ISCED 3).

The right-hand plot confirms that, as in the previous example, the communicative purpose does not relate strongly to difficulty.

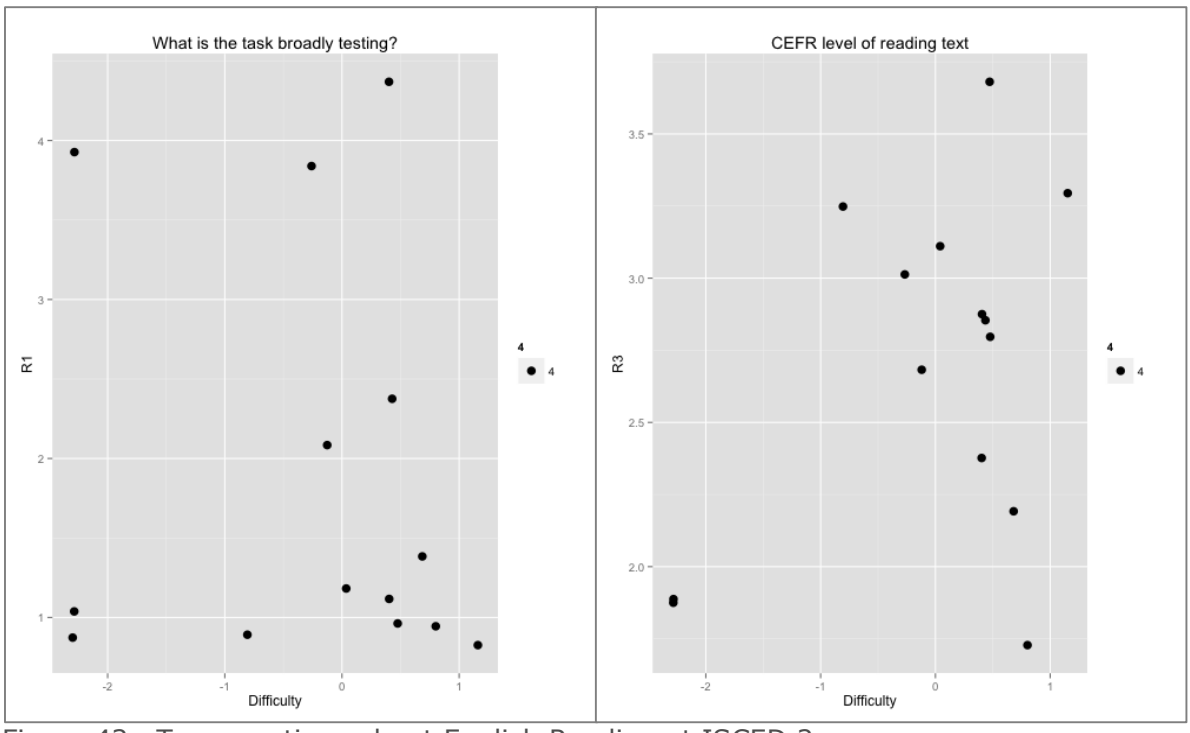


Figure 42 Two questions about English Reading at ISCED 2

The left-hand figure above illustrates two questions about English Reading. 'What is the task broadly testing?' offers a choice of four categories ordered as follows:

- careful reading - local
- careful reading - global
- expeditious reading - local
- expeditious reading - global.

The outcome here is curious because it appears that the tasks judged more difficult in the CJ exercise belong to the lowest category above - that of careful local reading. Is it possible that this indicates a validity problem? The progression intended by this scale relates to real-world reading objectives, with the final goal of fast efficient reading. It does not necessarily describe the world of the test, where high difficulty might be achieved by insisting on very careful reading in a way which does not reflect a real-world reading activity. Such considerations suggest the possibility that this kind of qualitative/quantitative comparative analysis might have diagnostic potential as a way of revealing problems of validity.

The right-hand figure, referring to the CEFR level of the text, behaves more as expected, with greater difficulty generally associated with higher levels.

The left-hand figure below shows that, as with the ISCED 2 task discussed above, the categories available for answering the question 'What is the task broadly testing?' do not seem to relate strongly to the difficulty of tasks as determined by the CJ exercise.

However, right-hand figure below the question as to the CEFR level of the reading text shows a loose but clear relationship with the difficulty of tasks as judged in the CJ exercise.

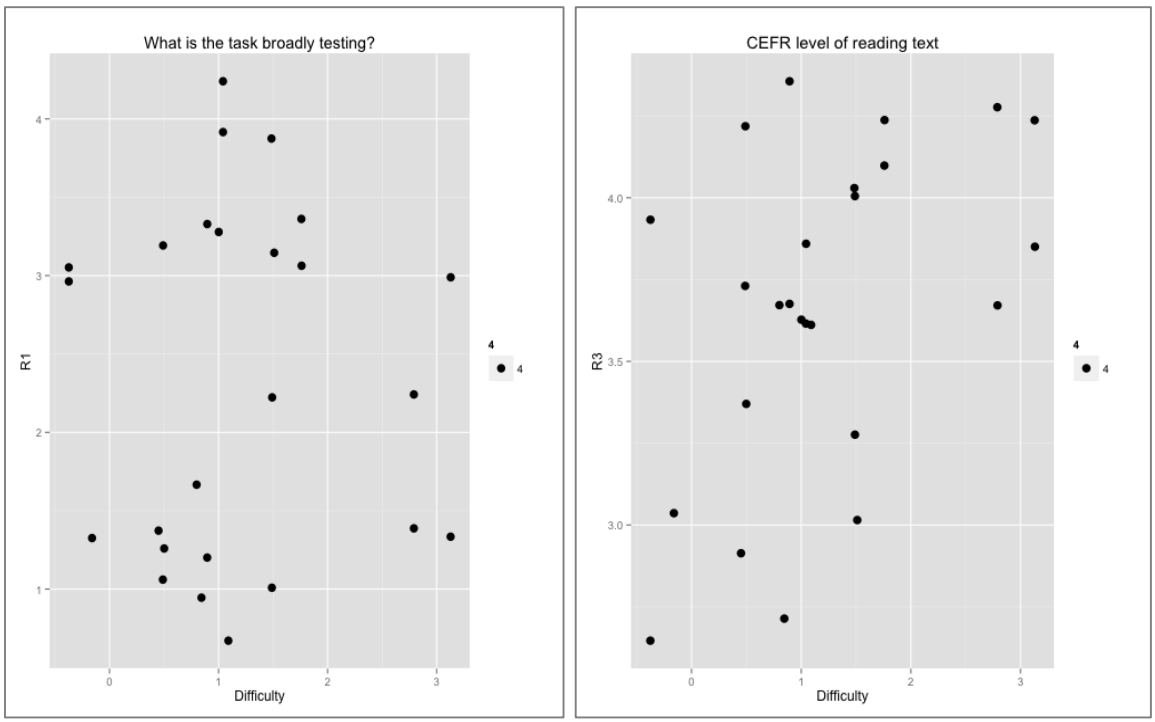


Figure 43 Two questions about English Reading at ISCED 3

This limited exploration of the relationship between the qualitative and quantitative data suggests that, as not unexpected, it is the clearly ordinal scales which will correlate best with the quantitative measures (though not necessarily agree with, in absolute terms). Thus the CEFR-related questions agree more closely than other categorically-expressed scales.

Overall, the comparison of the qualitative categories and the quantitative scale values indicates weak or non-existent correlation, at least for this limited sample of questions. It would appear that the descriptive categories used in the expert analysis do not necessarily relate strongly to task difficulty.

5.4 Comparative Judgement exercise: conclusions on comparability

Comparative Judgement appears to be a remarkably powerful tool for addressing the key goal of this project – to compare jurisdictions on the basis of different exam data. Due to the limited scope of this project, it was not possible to make full use of its potential, and in particular to compare performance samples or to link student performance to CEFR levels in the way that this has been demonstrated for test tasks. However, the potential of this technique for future comparability studies lies on the possibility of linking samples of students' performance from existing language examinations across EU Member States to the CEFR with limited efforts on the part of the national education authorities, as it was requested in the Council Conclusions in May 2014. Further considerations on how this methodology could be successfully applied in the future for the purposes of comparability are included in section 6 below.

5.4.1 The performance of judges

An aspect not yet addressed here concerns the performance of individual judges: some will of course be better than others (where "better" means agreeing more closely with other judges). A comparison of the distribution of infit statistics for the six tasks is shown below in Figure 44.

A high level of agreement between judges gives us more confidence in the outcomes. Infit is an IRT measure of how well a rater agrees with other raters. The infit statistic has an expected mean of 1, where lower values indicate that raters agree more than expected, and higher values indicate that raters agree less than expected, which may be more of a problem. We see that Reading and Writing for both English and French show a mean below or about 1. It is the two datasets comprising the reading and writing anchors which show a wide range of infit, and thus suggest more error.

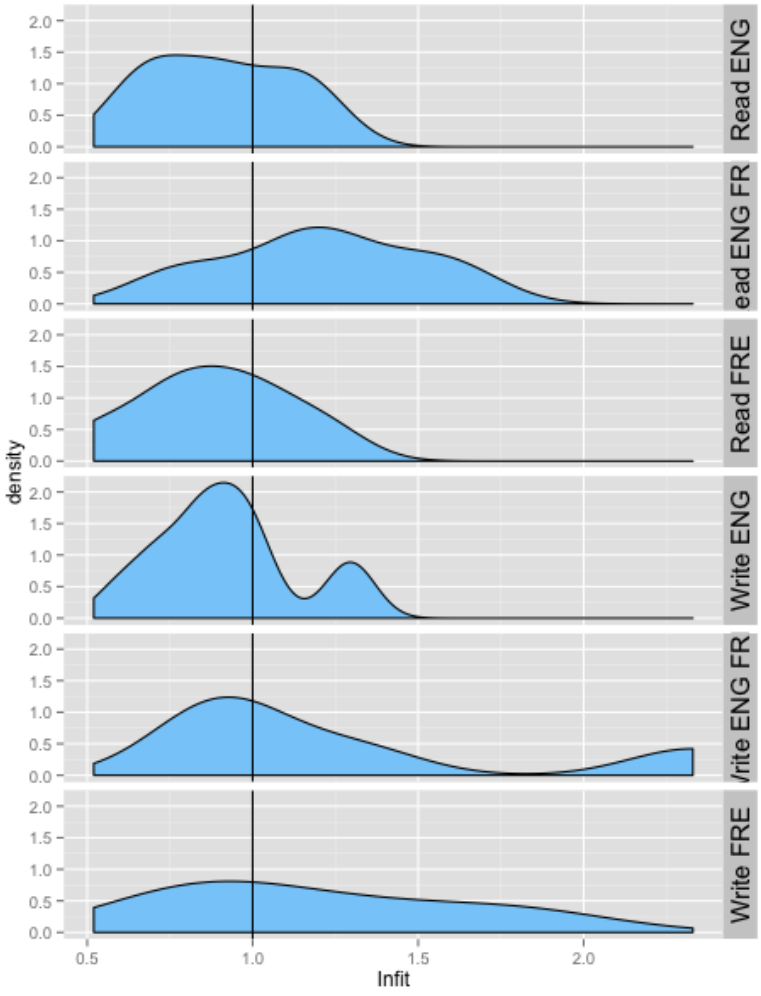


Figure 44 Distribution of judges' infit statistic for six comparative tasks

We must be cautious in intepreting this, given that the number of judged items is quite low for some of these groups, as shown in Table 11.

Table 11 Count of judged items in the six datasets and average judgements

	Count	Average number of Judgements
Read ANC	13	250
Read ENG	58	258
Read FR	10	242
Write ANC	7	236
Write ENG	19	398
Write FR	7	216

However, we might hypothesise on the basis of this limited data that the task of judging in two languages is intrinsically more difficult than judging in one language; even if we cannot be certain on this evidence.

5.4.2 The integrity of the measured traits

The standard error of the calibration of the test tasks is illustrated in the tables above. It can be seen to depend on the number of ratings and also on the point in the scale (precision is greatest in the middle of the scale), but in the case of English at least it appears very satisfactory.

5.4.3 The anchor to the European Survey on Language Competences

Having access to data from the first ESLC has proved very valuable. It has provided a reasonably valid link between the difficulty of jurisdictions' test tasks and the results of that survey.

6 Task 2: Proposals for ex-post adjustment to increase the comparability of existing national results

The question here is how to increase the comparability of existing results of national language tests. In order to answer this question, the first step is to look at the data available in each jurisdiction regarding the results of national language tests, and then to apply a methodology that would allow these national results to be mapped to a common framework of reference, enabling comparability of existing results across jurisdictions.

6.1 Proposed methodology

The methodology proposed for ex-post adjustment of existing results is illustrated in section 5.2. It involves the use of comparative judgements of test materials and, more importantly, of samples of students' performance for Writing and Speaking. This allows simultaneously giving each task or sample of performance a CEFR level and constructing a measurement scale. The scale is linked to the CEFR by seeding some tasks or samples with known difficulty levels. The result is that CEFR cut off points may be placed on the measurement scale and tasks and samples assigned a CEFR level. The next step, not completed in this study, is to link national performance results to the scale via the difficulty of the tasks used by a particular jurisdiction.

6.2 Issues related to this methodology

It is important at this point to highlight the difference between comparability of language examinations and comparability of national results stemming from these examinations. Two language tests may be comparable in terms of their measurement characteristics or on the basis of the constructs that they intend to measure. However, results of these two tests may still not be comparable if the format chosen to report these results differs too much from each other. In other words, results can only be compared if they report similar elements and in a similar format. The issue of the comparability of language tests has already been widely discussed in Task 1, and the following Task 3 puts forward proposals for development work to increase the comparability of existing language tests, which involves changing the nature of the exams and tests themselves according to some process of convergence. This section will therefore focus on the measures and procedures that could be implemented to increase the comparability of existing results of national language tests, and address the developments required to construct a comparative framework within which to address post-hoc adjustment and interpretation of current tests and exams.

6.3 Conditions for the successful application of this methodology

As shown in Task 1 and Task 5, there exist practical technical procedures capable of constructing the shared frame of reference which will be essential to any ex-post adjustment of national results. However, implementation of such procedures will be wholly dependent on well-coordinated international collaboration, the jurisdictions' commitment to provide all necessary data, and the need for an empowered, responsible and well-qualified body to implement the suggested procedures. Thus ex-post adjustment is not currently possible, but could become possible if sufficient attention was paid to it.

This will presumably only happen if jurisdictions can see benefit in setting up new systems to provide the necessary evidence. The benefit is real, if we accept the view of those who have criticised the current emphasis on standardised international educational surveys, which have been seen as distracting attention from more relevant educational research, focusing instead on headline comparisons of performance on an international league table. Any attempt to interpret and compare currently available evidence from exams and tests will inevitably focus attention on the fundamental issues – what are we teaching and how are we testing it? Does our use of exams promote learning or undermine it? What are reasonable targets in terms of learning outcomes? How efficient and effective is our teaching and learning, compared with neighbouring jurisdictions?

Before the suggested methodology for ex-post adjustment of existing national results can be fully implemented, there are three aspects that have to be addressed: the need for a common format for reporting national results across jurisdictions, the jurisdictions' commitment to provide all the data and test materials required in a suitable format, and the establishment of an annual schedule set and monitored by a responsible body.

6.3.1 A common approach to reporting national results

As evidenced in Task 5, jurisdictions are currently reporting national results in a wide range of formats which tend to only include those aspects relevant to their specific contexts. While the importance of reporting results in a way that is meaningful to its main stakeholders has to remain key, comparability of these results across jurisdictions will only be increased if these results are reported according to a common, pre-agreed set of straightforward instructions. This can be done instead or besides the reports that most jurisdictions already provide for national purposes, and it may as well be the case that the results in this specific format are only made available to the body in charge of monitoring the results across jurisdictions.

This responsible body will also have a reporting function as they will be in charge of determining the nature of the summary results provided for all jurisdictions. They might take an exclusively quantitative approach, effectively seeking to provide an annual 'league table' of jurisdictions, which will in turn determine the format in which jurisdictions may be required to provide the national results. It is also possible that, over time, the enterprise takes on a more formative function, which might prompt further work to study particular aspects of assessment and its impact on learning.

Where the focus of language education and assessment shifts towards effective language proficiency, then a critical issue to address becomes that of how to reconcile criterion-reference with the traditional and still important norm-referencing function of school assessment. Potential threats to language learning emerge. Some languages are harder than others: students or schools may discriminate against languages which bring them lower grades; English would become even more predominant. Strategies may need to be found so that all languages remain equally attractive to learners, or perhaps to make the lesser-spoken languages even more attractive.

6.3.2 Jurisdictions' commitment to provide relevant evidence

The form of evidence which jurisdictions found it easiest to provide seemingly relates to test materials. The analysis of these using qualitative and quantitative procedures has proved very informative, and the CJ exercise has demonstrated the practicality of

constructing a measurement scale based on such data. This is the first step in an evidential chain which would allow us to interpret students' performance on tests in terms of a common CEFR-linked scale.

The second step is the one for which jurisdictions have found it harder to provide evidence. This concerns the record of how students perform on tests and how this is interpreted.

The evidential chain to infer learning outcomes from test performance requires us to determine:

1. The difficulty of the test.
2. The performance level of each student on the test, i.e. their mark, either from right/wrong objective marking or from judgemental subjective marking.
3. How test difficulty and performance level are interpreted in terms of marks or grades.
4. How these grades are interpreted in terms of criterion-referenced levels of performance, such as the Common European Framework levels.

This is the fundamental data necessary in order to apply the method suggested in Task 1. It may well be that the data does not exist in explicit form, but can be extracted using psychometric models – we expect for example that relatively few jurisdictions would be aware of the difficulty of their tests, in measurement terms, because use of the necessary item-banking model, including pre-testing of items and their calibration, is not that widespread in educational testing (see Appendix 5 for an introduction to item banking).

The next issue, then, is how the data is to be analysed. The analysis suggested in this study has derived difficulties for the test tasks by using a Comparative Judgement method, and has related those difficulties to CEFR levels by including anchor tasks selected from the ESLC. The present study gives a prominent role to psychometric statistical procedures which are well established in assessment and increasingly highly developed. However, these are not familiar to much, or most, educational assessment in Europe, which tends to use more traditional approaches to quality, based largely on human judgement e.g. item-writers and markers deemed to be experts. Objective testing – that is, based on item responses which are summed to produce a percentage score – is widely practised, but generally without the psychometric measurement framework needed to support or validate it.

It is not our intention to discount the validity or reliability of testing done in traditional ways. In fact educational assessment of language competences still depends significantly on them, given that the performance skills of Speaking and Writing are still largely assessed using such subjective approaches. Direct judgement of performance skills has an immediacy, and a potentially closer link to learning, than indirect observation of objectively-marked skills such as Reading or Listening. Also the CJ approach given prominence in this study shows that human judgement can be effective and accurate in developing measurement scales, where channeled into the necessary psychometric procedures.

Performance examples of Writing or Speaking are another kind of evidence which we hoped to collect more of in the course of this study, but which jurisdictions found it difficult to supply. Given the very small number of samples of performance that we received, it would not have been worthwhile to include them in this study as they

would not have contributed towards the discussion about comparability. However, results of the analysis of samples of performance through CJ could yield very interesting and meaningful insights about how standards are understood and interpreted in each jurisdiction.

6.3.3 An annual schedule set and monitored by a responsible body

Clearly there will need to be excellent coordination across jurisdictions, facilitated by a responsible body. This body would have to agree on an annual schedule convenient to all jurisdictions (i.e. avoiding clashes with main examination periods or other busy periods in the year for national examination board), and establish a framework for planning the frequency of comparative studies, the groupings within which tests and examinations should be compared, etc. Furthermore, this body will also have to monitor the progress of these subsequent studies and suggest ways in which the procedures could be improved both across jurisdictions and at a national level to ensure the maximum comparability and representativeness of results.

7 Task 3: Proposals for development work to increase the comparability of existing language tests

For the purposes of the current section, the relevant findings reported in Task 1 may be divided into two groups. The first group of findings suggest things which all jurisdictions can consider ways to improve the quality of their tests. These include the selection and training of test constructors and raters (section 5.1.5 Personnel). For most exams, there was little evidence that those working on the tests were selected in accordance with relevant criteria and sufficiently trained after selection. If those involved in the process are not suitable for the role they are given, it is likely that results will include more measurement error, hence lower comparability. Improvement in these areas, however, often implies greater cost.

We should distinguish two groups of findings:

- findings which relate to improving the quality of the test in general, which should make for better comparability,
- findings which relate to diversity of constructs between tests - which is not a quality issue, but which will predictably decrease the comparability of test results.

Some test providers may decide that practical considerations outweigh those of quality or that some measures to improve quality are not feasible in view of factors such as the stakes of the test, the number of candidates taking the test and the budget available. An alternative, which can be noted at this point, and which has been partially implemented in Belgium and Portugal, is the use of tests provided by third parties. In these cases, economies of scale can make necessary quality enhancements possible.

The second group of findings should not be taken to imply that test providers need to change their tests. Different decisions may be made depending on the context of each test, considering factors such as the aims of the educational system, the content of the syllabus which accompanies the exams and the uses to which results are put. For example, according to European Commission/EACEA/Eurydice (2015), up to half of all national tests do not test all four skills and overall test results will therefore reflect different things. Another example is that of the domains within which topics for any content falls; for example, the occupational domain may be important in some contexts and not others and so will appear in some tests and not others and this will reduce comparability between test results.

If test providers were to change the fundamental characteristics of tests simply in order to make them more similar to those of another test provider, there would be a risk of losing qualities which are relevant to their own context.

We have pointed out that jurisdictions may have significantly differing intepretations of the CEFR levels, essentially norm-referencing them to their own levels of language-learning outcome. Counteracting this bias would at once improve comparability with no impact on the design of tests, and could be simply achieved by employing Comparative Judgement.

However, in light of the findings in section 5, it is possible to identify a number of proposals for development work that jurisdictions could undertake to increase comparability of existing language exams, and their quality at the same time. These

proposals are intended to increase the comparability of the constructs measured, the comparability of interpretations inferred from the tests, the comparability of the populations who take the test, and the comparability of the measurement characteristics in the exams.

7.1 Proposals to increase the comparability of constructs

The findings show that there is considerable diversity across the four CEFR levels targeted (A2-C1) by the tests under consideration for all the components (Speaking, Reading, Writing and Listening). The parameters examined are summarised in Table 12. In all cases, diversity which is likely to affect comparability was found. However, the context of use of each exam may have been an important influence in many of these cases. For this reason, a thorough review of the exam system, such as that provided by an ALTE audit (Saville, 2010), would be required to establish whether improvement could be gained by making changes based on these findings.

Table 12 Parameters examined for comparability of constructs

Skill	Parameter
speaking	interaction pattern
	communicative purpose
	domain
	rating criteria
writing	communicative purpose
	domain
	register
	rating criteria
reading	domain
	type of reading
	highest level of cognitive processing
listening	domain
	type of listening
	highest level of cognitive processing

Among the parameters which could be changed to improve the quality of the tests are communicative purpose, domain, type of reading/listening, and highest level of cognitive processing. This is because these parameters are expected to progress across CEFR levels. For communicative purpose, conative purpose (and perhaps

emotive purpose as well) would be found more frequently at higher levels, while phatic and referential purposes would be found at lower levels. This, to some extent, should follow domain, where personal might often co-occur with phatic communicative purposes, which would both occur at lower levels.

Reading or listening to long stretches of text at lower levels is difficult for learners as it places great demands on cognitive processing. Thus it is difficult to design items which address global or even expeditious reading/listening at lower levels. After this, however, it is likely that all four types of reading/listening would be important skills to develop and would be testable. For this reason, expanding the range of what is tested at B2 and above would be recommended, unless there were good reasons for not doing so. Finally, there is an expectation that higher levels of cognitive processing would be addressed at higher ability levels (Geranpayeh & Taylor, 2013; Khalifa & Weir, 2009). Thus for the skills of Reading and Listening, test constructors could articulate progression across cognitive stages, with a corresponding progression in difficulty attributable to contextual features, such as lexis or grammar. Such considerations could also be applied to the performance skills of Writing and Speaking.

Test providers might carry out initial research in order to determine whether any of the changes in any of the features discussed in this section would improve test quality.

7.2 Proposals to increase the comparability of interpretations

As noted above, Many test providers reference their interpretation to the CEFR, but the manner of doing this is not clear, and there is evidence that jurisdictions vary in their understanding of the levels, tending to norm them on their own context. None the less, it is our belief that the CEFR remains a viable and valid point of reference. The critical issue is: what do we mean by 'adopting the CEFR'?

The text of the CEFR betrays its multiple authorship, reflecting a range of influences:

- the functional/notional approach of Wilkins;
- needs-analysis, reflecting Trim's work on a unit-credit system for adult learners;
- the behavioural scaling descriptive approach of North's can-do scales;
- the action-oriented approach articulated by Coste.

Firstly, which of these should we adopt? Of course, the can-do scales provide a useful and complex picture of progression. There is only a problem if the scales are misinterpreted as the basis of a curriculum, which was never the intention. The action-oriented model strikes us as valuable and coherent with contemporary models of how learning happens. The *general competences* identified comprise:

- Knowledge, i.e. declarative knowledge (*savoir*)
- Skills and know-how (*savoir-faire*):
- Existential competence (*savoir-être*):
- Ability to learn (*savoir apprendre*): (Council of Europe 2001)

These are recognizably social-constructivist concepts, which allow us to see learning as a process which promotes the learner's personal development - of learning skills, attitudes and dispositions.

Secondly, what else must we add? Early critiques of the CEFR focused on the relatively poorly developed treatment of cognition. That has been addressed in work done by

Cambridge to develop constructs for the four skills of Listening, Speaking, Reading and Writing, based on Weir's (2005b) *socio-cognitive* validation model (Shaw and Weir (2007), Khalifa and Weir (2009) , Taylor (Ed) (2011) and Geranpayeh and Taylor (Eds) (2013)). Organised around Weir's validity model, these volumes set out to supply the useful level of detail which the descriptor scales of the CEFR itself do not. Concepts taken from these were used in the protocol for the qualitative analysis.

Corpus-based work within the English Profile (Hawkins & Filipovic, 2012) has contributed a further, linguistic dimension to the CEFR levels for English, and similar developments exist for other languages (See *www.coe.int/t/dg4/linguistic/DNR_EN.asp*).

Thus it is important to see the CEFR not as a closed set of prescriptions but rather as an area of ongoing development. Possible follow-ups to the present study might well contribute to that ongoing process.

For example, test results referenced to some other criterion-based framework might also be relatable to the CEFR. Challenges to comparability which this study highlights might become the focus of specific work. Tests which are currently norm-referenced might seek a basis in the psychometrics of this study to develop criterion-referenced interpretations; and so on.

Guidance is, of course, available on the Council of Europe's CEFR page http://www.coe.int/t/dg4/linguistic/default_en.asp?, and elsewhere)

7.3 Proposals to increase the comparability of populations

According to the ages of candidates, populations within each ISCED level were found to be relatively homogeneous. The age at which an ISCED test is taken is likely to be dictated by the educational system of the country. However, the collection of data on a wider range of candidate characteristics would be advisable, such as type of school, gender, years of target language instruction, etc. Such demographic data on test-takers gives a better view of efficiency and speed of learning, an issue which complicates making comparisons between the results of different tests.

7.4 Proposals to increase the comparability of tests' measurement characteristics

Measurement characteristics or conditions are features of the testing context which may decrease the comparability of measurement. Lack of training for personnel, for example, might lead to poorer execution of stages in the testing process (e.g. marking). This in turn will produce greater measurement error, making test results less reliable and therefore harder to compare across tests. Most recommendations in this section will results in improvements to the testing process, reduced measurement error and greater comparability. Such changes may, however, be expensive and must therefore be balanced with the practicality of producing the test.

Table 13 Parameters examined for comparability of measurement characteristics

Area	Parameter
personnel	selection of test constructors
	selection requirement for test constructors
	selection of markers and raters
	selection requirements for markers and raters
	training for markers and raters
	standardisation for markers and raters
procedures (performance testing)	number of raters per speaking performance
	number of raters per writing performance
procedures (standard setting and standard maintenance)	equivalent tests for all candidates in a single session
	equivalent tests for all candidates regardless of session
	approach to aligning the test to the CEFR
	targeting of CEFR levels
	estimation of reliability
materials (flaws)	in reading
	in listening
materials (context effects)	item types (linguistic competence)
	mode of prompt (speaking)
	control/guidance (writing)
	mode of input (writing)

Among the parameters investigated for the summary in Table 13 were those connected with the *recruitment and training of personnel*. Information was gathered about those involved in test construction generally as well as, more specifically, markers and raters. Language testing requires a range of expertise (e.g. test design, item writing, editing, marking or rating, analysis of results) but marking and rating

was focussed on because it is an area where individual variability can have a very large impact on the reliability and comparability of test results.. Little evidence was provided regarding procedures for reducing rater variability, so this is an area that should be reviewed. Practical steps to improve procedures is contained in ALTE & Council of Europe (2011). Appropriate expertise is required at each stage in the process.

Parameters relating to *performance testing* – particularly the number of ratings per performance for Writing and Speaking – are also important to reduce irrelevant variance. If more than one rater is used per performance, it is possible to compare the conduct of raters and derive a reliability index. If only one rating is given no such index can be derived and possible issues cannot be identified. Partial double marking, used operationally or in an ocasional research design, offers an effective solution. Approaches to this include sampling the work of raters to be checked by more experienced raters, seeding ready marked scripts, An operational procedure for resolving disagreements between raters is also recommended.

Another technique is to rely on the random distribution of scripts amongst raters and examine each rater's score profile. This may be effective where rather large numbers of scripts are involved. Test providers are recommended to ensure that adequate procedures for rating performance tests are in place.

Procedures relating to *standard setting and maintenance* bear directly on the comparability of tests. The requirement is to ensure that in each session tests vary only slightly in difficulty, and that grades can be placed taking into account any difference in difficulty which remains. For the skills of Reading and Listening item response theory (IRT) and an item banking approach to test construction, are the standard solutions. The difficulty of each task is known, allowing comparable tests to be constructed and precise grade cutoffs set. This process is described in more detail by North & Jones (2009). The Comparative Judgement approach used in this study offers a practical way of achieving the same end, avoiding the procedure of pretesting, which in many jurisdictions is deemed impractical. A CJ linking of the current test version to the previous test version can be done cheaply and securely, and with accuracy dependent on the number of raters involved.

Setting *CEFR-related standards* requires careful verification. The first requirement is that the test must be substantially coherent with the action-oriented, social-constructivist, socio-cognitive intention of the CEFR. Most current language tests should meet this requirement at least in part. The second requirement is to find a way of challenging the understanding of the CEFR levels in a given jurisdiction - it may well be specific to that jurisdiction. The use of sample test materials and performance exemplars is essential. The website CEFTrain (CEFTrain project 2015) is highly recommended in this regard.

Continuous improvement should be the goal of managing *overall test quality. Reliability* is an important property of a test which will benefit from (ALTE & Council of Europe (2011) quality management at different stages in the test construction process. Estimating reliability forms part of a diagnosis which can be the beginning of attempts to improve quality. Test providers should routinely collect response data (or score data from performance tests) and estimate reliability.

Flaws in test materials are due to flaws in the test construction process. Item writers may need better suppor, through training or item writer guidelines. Further iterative

stages of editing or pretesting could be introduced. Pretesting might be introduced where logistically or institutionally feasible. Analyses of data from pretesting help to detect flawed items. ALTE & Council of Europe (2011) has more information.

Many legitimate choices made when designing tests cannot be considered wrong but do, nevertheless, increase diversity between tests and decrease comparability between results. Examples of these features were also considered and the results show that there is some diversity between the tests examined. Many more such contextual effects could be examined, but results are likely to be similar and there is no obvious way to increase comparability because, based only on a cursory examination, all are equally appropriate. There is, therefore no recommendation for these features, other than for test providers to satisfy themselves that they are appropriate. If greater comparability is required, use of a single test for all jurisdictions would be the solution.

We have recognised that test constructs may vary, reflecting a jurisdiction's specific view of the purpose of language education. We expect however that where tests set out to measure a number of skills, at least some of these will be sufficiently close to a common core to enable useful comparison.

8 Task 4: Proposals for the development of future national language examinations

The recommendations contained in Task 3, based on this report's findings, are aimed at providers of existing tests. Most recommendations were aimed at improving test quality because without sufficient quality comparability between test results is impossible. Other findings showed simply that test providers may have made diverse but legitimate choices when designing their tests, which decrease comparability too.

Recommendations for jurisdictions planning to develop new tests are similar: quality is essential, for comparability, but also for interpreting and acting on test results. This section will focus on procedures with a relatively direct impact on our central concern with comparability; a more complete account of test quality is available in ALTE & Council of Europe (2011).

8.1 Recommendation 1: design the CEFR into the test

The task of designing comparable tests based on the CEFR will be easier if the CEFR is used as the starting point. Saying "based on the CEFR" we appeal to the notions discussed in 7.2 above. The extended system of description, illustration and practice which is now available to support the use of the CEFR is a valuable resource, certainly for the specific requirements of language testing, but also, we believe, for the successful integration of testing and language education.

As Milanovic (2009) makes clear, the CEFR is intentionally incomplete, so that it may be useful in a wide range of contexts. In relating a given context of learning or testing to the CEFR it is essential to remember that the CEFR is a frame of *reference* - the context is referred to the framework on its own terms: the framework is not imposed on the context

The implication of this is that it can only be a starting point for test providers, and they may need to develop the descriptions in order to apply them in their situation. A range of other tools are available, such as reference level descriptors (RDLs), which are compendiums of linguistic exponents at each CEFR level. The available RDLs can be found at *http://www.coe.int/t/dg4/linguistic/DNR_EN.asp*.

8.2 Recommendation 2: develop procedures to continually improve the test

ALTE & Council of Europe (2011) sees test provision as a cycle, where information is continually gathered in an attempt to detect issues and resolve them for future tests, if not the current test (see Figure 45). For example, the reliability of scores given by writing raters may be found to be low, and further investigation might identify the problem with interpretations of the ratiing scale. Improved or more regular training may be introduced. Continuing measures of rater reliability determine the impact of the new procedures, and whether more needs to be done. The data are also kept for later use. They can, for example, be exploited for more substantial revisions of the test, which may take place only periodically, or for the development of other tests.

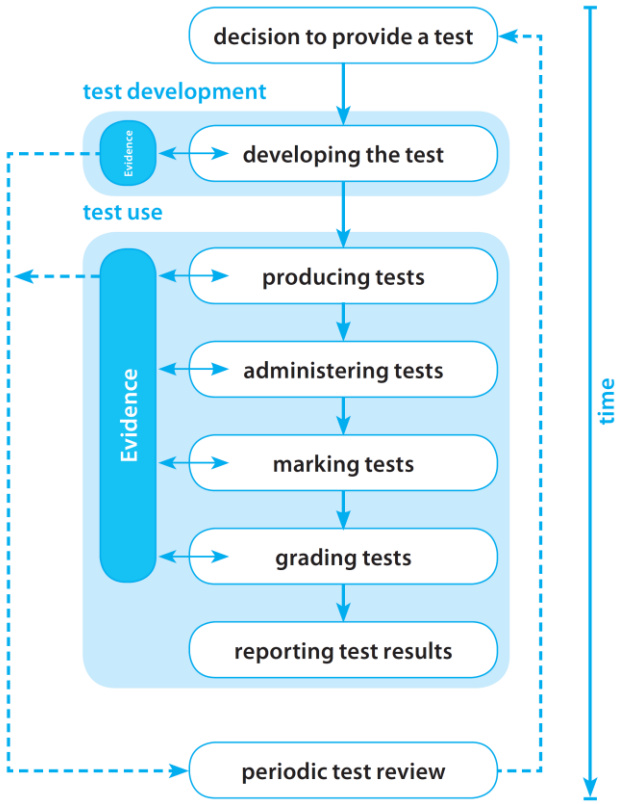


Figure 45 The testing cycle and periodic review (ALTE & Council of Europe, 2011:47)

8.3 Recommendation 3: develop a process to maintain standards

Ideally standard setting is done once only, because standards should not change over time. As tests inevitably vary in difficulty, the scores associated with standards may need to change from session to session. However, maintaining standards over time should not depend on human judgement. As described above, item response theory and an item banking approach to test construction enable standards to be maintained using psychometric methods. Comparative Judgement enables a similar approach to carrying a standard forward over sessions, and may be logistically simpler to manage. (See Appendix 5 for an introduction to Item Banking).

9 Task 5: Comparative overview of existing country data on language testing

This task required providing an overview of the data that is currently available from all jurisdictions regarding language test results. The focus of this task was only on results for the first foreign language in each jurisdiction, and the data should preferably come from publicly available sources.

9.1 Collection of national results

Out of the initial 133 language examiations included in this study, we attempted to collect data for 62 tests of first foreign languages from 33 jurisdictions, and we found that data was not available for 22 of these tests. This reduced the number of tests included in our summary to 40. There were different reasons why results may not be available and these are described in section 9.2 below. Some jurisdictions also reported results separately for all the different groups of students who would be taking the same test, which led to 5 additional results being added to our list, bringing the final number of national results examined to 45 (26 jurisdictions).

The data was collected by reviewing official documents which national education authorities usually make available on their websites. These documents and/or the information extracted from them was in most cases double-checked with the IEG members to make sure that the right data was being considered. In some other cases, the data was not publicly available but the IEG members were able to provide these details to the Core Project Team for the purposes of this study.

After the data was collected, the first attempt was to determine the scope of the data available for each examination and the type of information it contained. Data available differed greatly from jurisdiction to jurisdiction, and the aim of this approach was to identify the most common format of reporting national results across jurisdictions, which could be used in the future for a quicker extraction of national results directly from existing documents.

9.2 Availability and format of existing national results

There were a total of 25 tests (across 15 jurisdictions) for which relevant national results were not available, which was confirmed by IEG members from 11 jurisdictions. There were a number of reasons why data was not available, mainly due to lack of publicly available results or lack of official authorisation to disclose this information.

The format and presentation of the national results varied significantly among jurisdictions. While some jurisdictions provided reports that clearly summarised data, others provided raw figures from which it would be possible to calculate national results. After all the data were collected, the following observations were made regarding the current format in which national results of language tests are reported:

- Only 4 jurisdictions explicitly mentioned CEFR levels in their reports.
- The passing grade was publicly available for 35 tests (20 jurisdictions).
- No pass grade was found for 10 tests (7 jurisdictions).

- National results for 17 tests were reported by what percentage of students passed the test (1 jurisdictions).
- National results for 28 tests were reported by what percentage of students achieved each grade (20 jurisdictions).
- National results for 14 tests reported results for each skill (9 jurisdictions)
- The data for 23 jurisdictions (40 tests) clearly indicated the assessment scale against which results were reported.
- The data for 24 out of the 26 jurisdictions reported the number of students represented in the population but only half of these provided details of the types of candidates included in the sample.

9.3 Issues related to compiling a European summary table of adjusted national results

The collection and recording of national results showed an important diversity in the way these results are collected and reported across all the different jurisdictions. This diversity adds an extra layer of difficulty for comparability, especially if the ultimate aim is to produce a European summary table of adjusted national results which could be used to regularly monitor students' proficiency in one or several foreign languages. With this goal in mind, a number of factors need to be carefully considered to ensure if this summary table is to be compiled in the future, this is done in the most meaningful and representative way.

Through consideration of the **common themes in the national results**, it was determined that the most common way to report the national results data was by the percentage of students who passed each test. Whilst 10 jurisdictions explicitly reported on the percentage of students who passed, it was possible to find information in publicly available documents about the required pass grade for 20 other jurisdictions. Some jurisdictions only provided data about the percentage of students who attained each grade, whilst other jurisdictions gave results by the number of students, but this information was sufficient in many cases to calculate the percentage of students who had passed each exam on the basis of the pass grade for that examination.

The **use of "passing" grades**, which remains uninformative about the level of skills demonstrated, is on the one hand to be regretted, because it allows language education and testing to proceed without addressing the important question 'but what can they actually do?'. At the same time, it accommodates the need to set challenges matched to the current capacities of the candidature. That is, educational assessment is inevitably norm-referenced; the task for the future is to develop parallel criterion-referenced interpretations which focus the attention of students and teachers firmly on language proficiency.

It is also important to question how accurately the data represents the **test population**. Although jurisdictions specify whether the examinations test the whole population or only a sample of the students at that ISCED level, they do not generally record who is included in the summary data. For example, some jurisdictions clearly state that students with disabilities and adult learners are not included in the results, whilst others incorporate these students into the whole population. This makes it

difficult to determine whether the data accurately represents all students at a particular ISCED level.

Similarly, very few jurisdictions indicate whether **language tests are optional or compulsory**. If optional, then the outcomes should be considered to reflect a self-selected sample rather than the whole population. For example, students taking GSCE exams in England elect whether to take languages, and are thus not directly comparable with jurisdictions where language tests are compulsory for all students at that educational level (such as the *Überprüfung der Bildungsstandard* in Austria).

10 Conclusion

The idea of using existing exam data as a basis for making comparisons across jurisdictions is potentially of great significance. It would focus educational research in Europe less on comparisons based on international surveys and league tables, and more on the effectiveness of current teaching and assessment practices - which are, after all, the prerequisites for success in the wider world. In order to assess the feasibility of monitoring students' language competences through the results of language examinations implemented at a national level in the different EU Member States, this study aimed to complete five tasks as stated in section 2.4 above.

Task 1 (section 5) examined 133 language examinations from 33 jurisdictions (28 EU Member States), both qualitatively through expert content analysis and quantitatively through a Comparative Judgement exercise of Reading and Writing task difficulty. The analysis of this data shows that current national language examinations present a wide variety of features in terms of the constructs tested, the populations of test takers, the interpretations of the results and the measurement characteristics of these examinations. These features importantly determine test quality, and in turn impact on the validity and reliability of the results obtained. Furthermore, results of the Comparative Judgement exercise show that task difficulty for the same ISCED levels varies across jurisdictions. However, further data regarding students' scores and samples of performance would be needed to ascertain the CEFR level demonstrated by students in each jurisdiction.

Task 2 (section 6) argues for the application of the Comparative Judgement technique to a larger sample of tasks and, most importantly, to samples of students' performance in Writing and Speaking. This would allow for the ex-post adjustment of national results and their reliable mapping to the CEFR, which would therefore facilitate comparability of results across jurisdictions. In order for this technique to be fully effective, the importance of adopting a common approach to reporting results, the jurisdictions' commitment to provide relevant evidence, and setting up an annual schedule monitored by a responsible body were highlighted.

Task 3 (section 7) suggests that, on the basis of the findings in Task 1, a number of measures should be implemented by national examination boards in order to increase the quality of existing examinations, which would therefore make national results more valid and reliable, and which would in turn increase the potential for meaningful comparison of results across jurisdictions. Specific proposals were described in this section to help increase the comparability of the constructs tested, the interpretation of the results, the populations of test takers and the measurement characteristics of the tests, which would all have a positive effect on the quality of the examinations.

Task 4 (section 8) makes reference to a number of exisiting publications that could be useful when designing and implementing new national language examinations, and puts forward a number of recommendations that would especially increase the comparability of results of these exams in light of the findings in Task 1. These recommendations include designing the CEFR, developing procedures to continually improve the test, and developing a process to maintain standards.

Task 5 (section 9) includes an overview of the data that is currently available from all jurisdictions regarding national test results for the first foreign language in each jurisdiction. From the 45 examinations (26 jurisdictions) for which relevant data was

found, it emerged that national results currently being reported vary greatly among jurisdictions both in content and format. The compilation of a European summary table of national results would require therefore, as a first step, some agreement on the part of the different educational authorities about the data which must be reported and the format in which this data must be provided. A number of considerations for the meaningful compilation and interpretation of such a table are also provided, and include the selection of the data that is to be reported, the meaning of "passing" grades, and the test populations.

The findings from this study confirm therefore that the assessments produced by jurisidictions differ over a range of features, in matters of detail and of substance. In other words, the challenge is that national results are not only reported in a wide range of formats – which on its own makes it already difficult to compare results among jurisdictions – but, most importantly, that the language tests themselves show great diversity in the understanding of the contructs tested, the interpretations to be inferred from the results, the populations who take the exams and the measurement characteristics of the tests.

It is important to distinguish differences motivated by choice of educational objectives, which should be respected, and differences which are not essential and could readily be reduced or eliminated through forms of coordinated action. If the primary goal is to achieve comparability of language learning standards across jurisdictions, then addressing two fundamental issues would signficantly improve the situation: the attribution of CEFR levels, and the lack of control over standards across exam sessions.

Both of these could be relatively easily achieved: exemplars of performance skills could be shared across all jurisdictions, providing a common framework of reference. Control over standards across sessions would ideally involve more sophisticated psychometric interventions. While pretesting may be deemed impractical by many jurisdictions, the Comparative Judgement approach illustrated in this study offers a practical and potentially effective solution, especially if samples of Writing and Speaking performance were made available.

Two important questions are therefore raised by this study and should be given careful consideration before any further steps are taken towards increasing comparability of national language tests:

- to what extent are jurisdictions willing and able to improve current approaches?
- to what extent is some higher level of collaboration between countries necessary or possible?

The conduct and outcomes of this study illustrate that comparability is not simply a psychometric issue, but depends greatly on documentation and forms of collaboration. Relevant evidence regarding the different national language examinations is therefore essential to allow for comparability of these examinations. The importance of developing comparable forms of documentation needs to be highlighted, and it would be a valuable first objective if jurisdictions wish to follow up on the outcomes of this study. However, **the meaningful comparability of national results of language examinations across EU Member States will not only depend on these results being expressed in a uniform format, but also on implementing measures at both national and European level that would increase the quality of current**

language examinations and ensure results are similary valid and reliable across all jurisdictions.

11 References

Alderson, J. C., Figueras N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2004). The development of specifications for item development and classification within the Common European Framework of Reference for Languages: learning, teaching, assessment. Reading and listening. Final report of the Dutch CEF construct project. Unpublished document.

Alderson, J. C., Figueras N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project, *Language Assessment Quartery* 3 (1), 3-30.

ALTE, & Council of Europe. (2011). *Manual for Language Test Development and Examining for use with the CEFR.* Retrieved from http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf

Andrich, D (1978) Relationships between the Thurstone and Rasch approaches to item Scaling, *Applied Psychological Measurement* 2, 449–460.

Bramley, T. (2005) A Rank-Ordering Method for Equating Tests by Expert Judgment, *Journal of Applied Measurement 6/2, 202–223*.

Breton, G. (2008). *Cross-language benchmarking seminar to calibrate examples of spoken production in English, French, German, Italian and Spanish with regard to the six levels of the Common European Framework of Reference for Languages (CEFR)*. CIEP, Sèvres, 23-25 June 2008 REPORT

Council of Europe. (2001). 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 (CEFR) - A Manual. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf

European Commission. (2014). *The Structure of the European Education Systems 2014/15: Schematic Diagrams.* schematic diagrams can be found at: http://eacea.ec.europa.eu/education/eurydice/facts_and_figures_en.php#diagrams

European Commission. (2015). *Proposed Mandate, Indicator Expert Group on Multilingualism*. Strasbourg: European Commission.

European Commission/EACEA/Eurydice. (2015). *Languages in Secondary Education: An Overview of National Tests in Europe - 2014/15*. Eurydice Report. Luxembourg: Publications Office of the European Union.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), Examining Listening: Research and Practice in Assessing Second Language Listening (Studies in Language Testing Vol. 35). (pp.77-151). Cambridge: UCLES/Cambridge University Press.

Geranpayeh, A., & Taylor, L. (2013). Examining Listening: Research and Practice in Assessing Second Language Listening (Studies in Language Testing Vol. 35). Cambridge: UCLES/Cambridge University Press.

Green, A. (2012). Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition Across the Ability Range. (English Profile Studies Vol. 2). Cambridge: UCLES/Cambridge University Press.

Jones (2002) Relating the ALTE framework to the Common European Framework of Reference, in Alderson, J C (Ed.),167–83.

Jones (2005) Raising the Languages Ladder: constructing a new framework for accrediting foreign language skills, *Research Notes* 19, 15–19, Cambridge: Cambridge ESOL.

Jones (2009b) A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard setting, *Research Notes* 37, 6–9, Cambridge: Cambridge ESOL.

Jones, N and Saville, N (2009) European Language Policy: Assessment, Learning and the CEFR, *Annual Review of Applied Linguistics*, 29, 51–63.

Jones, N. & Saville, N. (2007). Scales and frameworks. In Spolsky, B. & Hult, F. M. (Eds) The Handbook of Educational Linguistics. (495-509). (London: Wiley-Blackwell.

Jones, N. (2014). Multilingual Frameworks: The Construction and Use of Multilingual Proficiency Frameworks (Studies in Language Testing Vol.40). Cambridge: UCLES/Cambridge University Press.

Khalifa, H., & Weir, C. J. (2009). Examining Reading: Research and Practice in Assessing Second Language Reading (Studies in Language Testing Vol. 29). Cambridge: UCLES/Cambridge University Press.

Kolen, M. J., & Brennan, R. L. (2004). Test Equating, Scaling, and Linking: Methods and Practices (2nd ed.). USA: Springer.

Linacre, J M (2006) Rasch Analysis of Rank-Ordered Data, *Journal of Applied Measurement*, 7/11, 129–139.

Milanovic, M. (2009). Cambridge ESOL and the CEFR. Research Notes, 37, 2-5.

North, B. J., & Jones, N. (2009). Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) - Further Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgement and IRT Scaling Retrieved from http://www.coe.int/T/DG4/Linguistic/Manual%20-%20Extra%20Material%20-%20proofread%20-%20FINAL.pdf

Reckase, M. D. (2009). Standard setting theory and practice: issues and difficulties. In N. Figueras & J. Noijons (Eds.), Linking to the CEFR levels: Research perspectives. (pp.13-20). Arnhem: CITO/EALTA.

Shaw, S. D., & Weir, C. J. (2007). Examining Writing: Research and Practice in Assessing Second Language Writing (Studies in Language Testing Vol. 26). Cambridge: Cambridge University Press.

Taylor, L. (Ed.). (2011). Examining Speaking: Research and Practice in Assessing second Language Speaking (Studies in Language Testing Vol. 30). Cambridge: Cambridge University Press.

Thurstone, L L (1927) A law of comparative judgment, *Psychological Review* 3, 273–86.

University of Cambridge ESOL Examinations. (2011). Principles of Good Practice - Quality Management and Validation in Language Assessment. Cambridge: University of Cambridge ESOL Examinations.

Weir, C. J. (2005). Language Testing and Validation: An Evidence-based Approach. Basingstoke: Palgrave Macmillan

Appendix 1 List of exams included in the study

Country	ISCED Level	Exam Name	Included	Reasons for inclusion/ non inclusion
Austria	ISCED 2	Überprüfung der Bildungsstandards Assessment of National Education Standards	Yes	
Austria	ISCED 3	Standardisierte kompetenzorientierte Reifeprüfung Academic Secondary Schools (AHS) Standardised and Competence-oriented Matriculation Examination	Yes	
Belgium FR	ISCED 2	Certificat d'études du premier degré de l'enseignement secondaire Certificate of First Stage Studies	Yes	
Belgium FR	ISCED 3	There is no exam at this level in Belgium FR		
Belgium GE	ISCED 2	Nachweis grundlegender Kenntnisse in der Französischen Sprache, Niveau B1 Certificate in French Language Studies B1	Yes	
Belgium GE	ISCED 3	Nachweis grundlegender Kenntnisse in der Französischen Sprache, Niveau B2 Certificate in French language Studies B2	Yes	
Belgium NL	ISCED 2	Peiling Frans in de eerste grad secundair onderwijs (A-strom): lezen, luisteren en schrijven National assessment in the first year of secondary education (A-strom): reading, listening and writing	Yes	
Belgium NL	ISCED 3	Peiling Frans luisteren en spreken in de derde grad aso, kso en tso van het secundair onderwijs	Yes	

Country	ISCED Level	Exam Name	Included	Reasons for inclusion/ non inclusion
		National assessment in the third year "aso", "kso" and "tso" of secondary education		
Bulgaria	ISCED 2	Национално външно оценяване National External Examination	Yes	
Bulgaria	ISCED 3	национално външно оценяване матура National External Evaluation / Matriculation	Yes	
Croatia	ISCED 2	There is no exam at this level in Croatia		
Croatia	ISCED 3	Ispit državne mature State Matriculation Exam	Yes	
Cyprus	ISCED 2	There is no exam at this level in Cyprus		
Cyprus	ISCED 3	Παγκύπριες Εξετάσεις Pancyprian Examinations	Yes	
Czech Rep	ISCED 2	Výběrové ověřování výsledků žáků na úrovni 4. a 8. ročníků základních škol a 2. ročníků středních odborných škol Sample survey on pupils' results in the 4th and 8th grade of basic schools and pupils in the 2nd years of upper secondary vocational schools	Yes	
Czech Rep	ISCED 3	Výběrové ověřování výsledků žáků na úrovni 4. a 8. ročníků základních škol a 2. ročníků středních odborných škol Sample survey on pupils' results in the 4th and 8th grade of basic schools and pupils in the 2nd years of upper secondary vocational schools	Yes	
Denmark	ISCED 2	Folkeskolens prøver Folkeskole Leaving Examination (Form 9)	Yes	

Country	ISCED Level	Exam Name	Included	Reasons for inclusion/ non inclusion
Denmark	ISCED 2	Folkeskolens 10. Klasseprøver Folkeskole Leaving Examination (Form 10)	Not included	Optional exam
Denmark	ISCED 3	General Upper Secondary School Examination (STX)	Yes	Exam taken by the largest proportion of students at this ISCED level
Denmark	ISCED 3	Higher Commercial Examination (HHX)	Not included	Exam taken by a small proportion of students at this ISCED level
Denmark	ISCED 3	Higher Technical Examination (HTX)	Not included	Exam taken by a small proportion of students at this ISCED level
Denmark	ISCED 3	Higher Preparatory examination (HF)	Not included	Exam taken by a small proportion of students at this ISCED level
Denmark	ISCED 3	EUX programme, a combination of a general and vocational upper secondary education	Not included	Exam taken by a small proportion of students at this ISCED level
Estonia	ISCED 2	Riiklik tasemetöö National standardized test	Not included	Not yet implemented. No date of implementation fixed
Estonia	ISCED 3	Gümnaasiumi lõpueksamid External school leaving examinations	Yes	

Country	ISCED Level	Exam Name	Included	Reasons for inclusion/ non inclusion
Finland	ISCED 2	Toinen kotimainen kieli, ruotsi B-kielenä 9. Vuosiluokalla, Toinen kotimainen kieli, suomi A-kielenä ja äidinkielenomainen suomi 9. Vuosiluokalla Second National Language: Swedish as B language, grade 9; Finnish as A-language; or Native Level, grade 9	Not included	Not sufficient information available
Finland	ISCED 2	Vieraat kielet 9. Vuosiluokalla (Foreign languages in grade 9)	Not included	Not sufficient information available
Finland	ISCED 3	Ylioppilastutkinto Matriculation Examination (General Upper Secondary Education only)	Yes	
France	ISCED 2	Cycle des évaluations disciplinaires réalisées sur échantillons CEDRE Assessment (Grade 9 sample-based subject evaluations)	Yes	
France	ISCED 3	Baccalauréat général et technologique General and Technological Baccalaureate	Yes	
Germany	ISCED 2	Ländervergleich Sprachen National Assessment Study	Yes	
Germany	ISCED 2	Vergleichsarbeiten – comparison tests VERA Comparison Tests	Yes	
Germany	ISCED 3	There is no exam at this level in Germany		
Greece	ISCED 2	There is no exam at this level in Greece		
Greece	ISCED 3	Ειδικό Μάθημα Ξένης Γλώσσας Modern Foreign Languages Examination	Yes	

Country	ISCED Level	Exam Name	Included	Reasons for inclusion/ non inclusion
Hungary	ISCED 2	Idegen nyelvi mérés Foreign Language Assessment (grade 6)	Not included	Exam at the end of relevant ISCED level preferred
Hungary	ISCED 2	Idegen nyelvi mérés Foreign Language Assessment (grade 8)	Yes	
Hungary	ISCED 2	Célnyelvi mérés Target language assessment grade 10 (end of compulsory education)	Not included	Exam taken by a small proportion of students at this ISCED level
Hungary	ISCED 2	Célnyelvi mérés Target language assessment grade 8 (end of lower secondary education)	Not included	Exam taken by a small proportion of students at this ISCED level
Hungary	ISCED 3	Érettségi Upper Secondary School Leaving Examination	Yes	
Ireland	ISCED 2	Junior Certificate Examination	Yes	
Ireland	ISCED 3	Leaving Certificate Examination	Yes	
Italy	ISCED 2	There is no exam at this level in Italy		
Italy	ISCED 3	Seconda prova scritta dell'Esame di Stato National Examinations: Second Written Paper	Yes	
Latvia	ISCED 2	Eksāmens svešvalodā 9.klasei (angļu, vācu, krievu, franču val.) Grade 9 Foreign Language Examination	Yes	
Latvia	ISCED 3	Centralizētais eksāmens par vispārējās vidējās izglītības apguvi svešvalodā Centralised Secondary School Leaving Examination in Foreign Languages	Yes	

Country	ISCED Level	Exam Name	Included	Reasons for inclusion/ non inclusion
Lithuania	ISCED 2	Užsienio kalbos lygio nustatymo testas Foreign Language Standardised Test	Yes	
Lithuania	ISCED 3	Užsienio kalbos valstybinis brandos egzaminas State Matriculation Examination in Foreign Languages	Yes	
Lithuania	ISCED 3	Užsienio kalbos (anglų, prancūzų, rusų,vokiečių) įskaita Foreign Language Speaking Credit	Yes	
Luxembourg	ISCED 2	Epreuves Standardisées Standardised Tests	Yes	
Luxembourg	ISCED 2	Epreuves Communes Français/ Allemand National Tests in French/ German	Yes	
Luxembourg	ISCED 2	Epreuves Communes Anglais National Tests in English	Yes	
Luxembourg	ISCED 3	There is no exam at this level in Luxembourg		
Malta	ISCED 2	Annual Examinations for Secondary Schools (Form 1)	Not included	Exam at the end of relevant ISCED level preferred
Malta	ISCED 2	Annual Examinations for Secondary Schools (Form 2)	Not included	Exam at the end of relevant ISCED level preferred
Malta	ISCED 2	Annual Examinations for Secondary Schools (Form 3)	Not included	Exam at the end of relevant ISCED level preferred
Malta	ISCED 2	Annual Examinations for Secondary Schools (Form 4)	Not included	Exam at the end of relevant ISCED level preferred

Country	ISCED Level	Exam Name	Included	Reasons for inclusion/ non inclusion
Malta	ISCED 2	Annual Examinations for Secondary Schools (Form 5)	Yes	
Malta	ISCED 2	MATSEC Secondary Education Certificate	Not included	Another exam at the same ISCED level and grade was preferred.
Malta	ISCED 3	MATSEC Matriculation Certificate	Yes	
Netherlands	ISCED 2	Centraal examen VMBO National Examination VMBO	Yes	
Netherlands	ISCED 3	Centraal examen HAVO National Examination HAVO	Yes	
Netherlands	ISCED 3	Centraal examen VWO National Examination VWO	Yes	
Poland	ISCED 2	Egzamin gimnazjalny z języka obcego nowożytnego (poziom podstawowy lub poziom rozszerzony) End of Lower Secondary Education Language Examination (Basic and Extended Level)	Yes	
Poland	ISCED 3	Egzamin maturalny z języka obcego nowożytnego (poziom podstawowy, poziom rozszerzony lub poziom dwujęzyczny) Matriculation Language Examination (Basic, Extended or Bilingual level)	Yes	
Portugal	ISCED 2	Key English Test for Schools (KETfS)	Yes	
Portugal	ISCED 3	Exame Final Nacional do Ensino Secundário National Secondary Education Final Test	Yes	

Country	ISCED Level	Exam Name	Included	Reasons for inclusion/ non inclusion
Romania	ISCED 2	Evaluarea naţională la finalul clasei a VI a în aria curriculară "Limbă și Comunicare" - limba română și o limbă străină National Evaluation in the Language and Communication Curriculum Area	Yes	
Romania	ISCED 3	Examenul de bacalaureat Proba C de evaluare a competenţelor lingvistice într-o limbă de circulaţie internaţională studiată pe parcursul învăţământului liceal Test C of the National Baccalaureate Examination: Grades XII and XIII	Yes	
Slovakia	ISCED 2	There is no exam at this level in Slovakia		
Slovakia	ISCED 3	Externá časť maturitnej skúšky a písomná forma internej časti maturitnej skúšky School Leaving Examination: External and Internal Written Parts	Yes	
Slovenia	ISCED 2	Nacionalno preverjanje znanja (NPZ) National Assessment of Knowledge	Yes	
Slovenia	ISCED 3	Splošna matura General Matriculation Examination	Yes	
Slovenia	ISCED 3	Poklicna matura* Vocational Matriculation Examination	Yes	
Spain (Navarre)	ISCED 2	Evaluación diagnóstica censal 2° de ESO, Competencia lingüística en inglés Diagnostic evaluation, 2nd year, Linguistic competence in English	Yes	

Country	ISCED Level	Exam Name	Included	Reasons for inclusion/ non inclusion
Spain (Catalonia)	ISCED 2	Avaluació educació secundària obligatòria 4t d'ESO, Competència lingüística: llengua anglesa Evaluation in compulsory secondary education, 4th Year, Linguistic competence: English	Yes	
Sweden	ISCED 2	Nationellt prov National Test	Yes	
Sweden	ISCED 3	Nationellt prov National Test	Yes	
UK England	ISCED 2	(General Certificate of Secondary Education (GCSE)	Yes	
UK England	ISCED 3	Advanced Subsidiary Level (AS Level)	Not included	Exam at the end of relevant ISCED level preferred
UK England	ISCED 3	General Certificate of Education Advanced Level (GCE A Level, or A Level – A2)	Yes	
UK Northern Ireland	ISCED 2	(General Certificate of Secondary Education (GCSE)	Yes	
UK Northern Ireland	ISCED 3	Advanced Subsidiary Level (AS Level)	Not included	Exam at the end of relevant ISCED level preferred
UK Northern Ireland	ISCED 3	General Certificate of Education Advanced Level (GCE A Level, or A Level – A2)	Yes	
UK Scotland	ISCED 2	Nàiseanta 5 National 5	Yes	
UK Scotland	ISCED 3	Àrd-ìre Highers	Yes	
UK Scotland	ISCED 3	Àrd-ìre Adhartach Advanced Higher	Not included	Not yet implemented
UK Wales	ISCED 2	(General Certificate of Secondary Education (GCSE)	Yes	

Country	ISCED Level	Exam Name	Included	Reasons for inclusion/ non inclusion
UK Wales	ISCED 3	Advanced Subsidiary Level (AS Level)	Not included	Exam at the end of relevant ISCED level preferred
UK Wales	ISCED 3	General Certificate of Education Advanced Level (GCE A Level, or A Level – A2)	Yes	

Appendix 2 The content analysis tool

Introduction

The purpose of this protocol is to ensure coherent and comparable description of countries' tests or exams. The available data come from several sources:

- A recent Eurydice survey of jurisdictions capturing basic parameters of tests
- Examples examples of test or exams provided by the countries
- Possiblypossibly, curricular statements or other documentation
- possibly, performance samples for Speaking and Writing.

NB there may be insufficient evidence to answer some questions adequately, or at all. Our agreement with the European Commission is that we can go back to countries for them to check our analysis and to provide specific further information; but we have been asked to use existing data as far as possible, and not to use a country questionnaire.

We wish to construct a qualitative picture of a country's language education goals, and how/whether these are reflected in the test design. Please attempt this based on examination of test design and curricular documents; we will then check with or seek further information from countries.

Analyst's name

- The exam or test: high-level description: purpose and design

Pages 2 to 7 concern the overall design and purpose of the test, and features relating to its validity and reliability.

(Sources: ALTE CEFR grids for Writing, Speaking + other)

- Design and purpose

1. Country
2. Whole cohort or a sample? (Eurydice information)
    a. Whole cohort
    b. Sample
3. Date of last revision (Eurydice information)
4. Name of test (Eurydice information)
5. ISCED level
    a. ISCED 2
    b. ISCED 3
6. Total exam duration
7. Frequency of administration (Eurydice information)
8. If sample, what evidence is provided of the samples being representative?

9. Is the test paper-based or computer-based? (Eurydice information)
    a. Paper-based
    b. Computer-based

Component	Number of items per component	Time allocation per component (in minutes)	Weighting of component in total (in percentage)
Reading			
Listening			
Writing			
Speaking			
Structural competence			
Other			

10. Total exam duration (in minutes)
11. Test purpose (check all that apply)
    a. Achievement of curricular objectives
    b. Diagnosis (individual)
    c. Monitoring (population)
    d. Language proficiency, criterion-referenced to e.g. CEFR
    e. Other (please specify)
12. Test purpose – If for this given test the official purposes are more than one and are not given the same priority, please state clearly what the prime official purpose(s) is (are).
13. Test purpose- Please state any accountability issues, particularly with respect to school leaving exams (who is impacted, what is likely effect on teaching to test, etc.)
14. Language of instructions
    a. Target language
    b. Students' native language
    c. Other
15. Language level of task instructions
    a. Below target level
    b. Same as target level
    c. Above target level
16. Control/guidance by task rubric
    a. Rigid
    b. Open format

- Goals of language education

Please rank the importance of these goals – in teaching and in testing – on a scale from 1 (highest) to 7 (lowest)

Which goals are most important in language education – is this reflected in the exam?

Goals	Importance in teaching	Importance in testing
Social		
Academic		
Professional		
Mathetic (as a tool for learning)		
Imaginative		
Literary		
Intercultural competence		

Comments

Which language activities are most important in language education – is this reflected in the exam?

Language activities	Importance in teaching	Importance in testing
Listening		
Oral production		
Oral interaction		
Mediation		
Reading		
Written production,		
Written Interaction		

Comments

To summarise (please consult documentation):

	Very poor	Poor	Adequate	Good	Very good
Adequacy of evidence for judging this					
Overall judgement on quality of this					

Feedback

24. Is feedback provided to students?
25. If so, is feedback: (please check all that apply)
    a. Criterion-related: (e.g. CEFR level),
    b. Quantitative (percentage or raw score, test-specific, grade or ranking?)
    c. Qualitative evaluation?
        - Reliability

This section concerns several aspects of reliability

26. Are tests for a given session standardised (i.e. are they the same for all students in the jurisdiction)?
27. Are tests standardised across sessions (i.e., is the difficulty of this year's test demonstrably the same as the difficulty of last year's test)?
28. If so, how is this achieved?
29. Are reliability indices estimated:
    a. for students' test scores,
    b. for the performance (accuracy) of raters?
30. To summarise: (please consult documentation)

	Very poor	Poor	Adequate	Good	Very good
Adequacy of evidence for judging this					
Overall judgement on quality of this					

Interpretation

Is test performance interpreted in terms of the CEFR or in other ways?

31. Is the CEFR used to rate students' performance in this test? (Eurydice information)
32. If so, please indicate the CEFR level(s) tested for this test. (Eurydice information)
    a. A1

b. A2
c. B1
d. B2
e. C1
f. C2

33. Is the test aligned to the Common European Framework (CEFR)?
34. If so, how was this alignment established? (Does it depend uniquely on can-do statements or any other approach? What references are cited?)
35. What evidence for it exists?
36. If the test is not aligned to the CEFR: how are test results interpreted? Do scores or rating scales reflect similar criterion-referenced levels? If so, please estimate the correspondence to the CEFR.
37. And how have these interpretations been developed?
38. To summarise: (please consult documentation)

	Very poor	Poor	Adequate	Good	Very good
Adequacy of evidence for judging this					
Overall judgement on quality of this					

Future reforms on national tests in languages

This information can be extracted from the data provided by Eurydice, Question 5. If there are no reforms planned, please leave blank and click 'Next' to the following page.

39. Do planned reforms include…
    a. Changes which address the construct tested?
    b. Changes which address the reliability of assessment?
    c. Changes which modify the purpose of the assessment?
40. In which way are planned changes going to affect each of the above three categories?

Test construction, marking and grading

This page concerns the training of those involved in test conduct.

41. Who designs the tests? (Eurydice information)
    a. Teachers
    b. Researchers
    c. Inspectors
    d. A team of the above
    e. Other (please specify)
42. What criteria are used for selection and training of test constructors?
43. Who oversees the test? (Eurydice information)
    a. People working inside the school
    b. People working outside the school

c. Both
d. Other (please specify)

44. What criteria are used for the selection and training of markers and raters?
45. What criteria are used for selection and training of those responsible for standard setting and maintaining?
46. Who scores the test? (Eurydice information)
    a. Electronically scored
    b. Teachers working inside the school
    c. Teachers working outside the school
    d. Other (please specify)
47. To summarise: (please consult documentation)

	Very poor	Poor	Adequate	Good	Very good
Adequacy of evidence for judging this					
Overall judgement on quality of this					

- Speaking

48. What is the first foreign language that have you been asked to look at for this test?
    a. English
    b. French
    c. Spanish
    d. German
    e. Italian

Speaking: rating

Rating method, rating criteria

49. Is the topic known to students in advance, so that performance is rehearsed?
50. Rating method is:
- holistic
- analytic: band descriptors
- analytic: checklist
- Other

51. Rating criteria (check all that apply)
- argumentation
- interactive communication
- grammatical accuracy and/or range
- lexical accuracy and/or range
- pronunciation
- other (please specify)

52. How many raters per performance?
53. Is there a procedure in case of disagreement?

54. Criteria are known to student

Speaking: tasks

Please complete this grid for each task, to a maximum of five tasks

55. Speaking tasks

Prompt

- audio/video
- oral only (by examiner)
- picture
- text

Interaction

- with examiner
- with other test-taker(s)
- with recorded prompt
- monologue (no interaction)

Response type

- short monologue
- extended monologue
- short interaction
- extended interaction

Domain

- personal
- public
- occupational
- educational/academic

Integration with other skills

- none
- reading
- writing
- listening

Communicative purpose

- referential (telling)
- emotive (reacting)
- conative (argumentation, persuasion)
- phatic (social interaction)

Expected level of response

- A1
- A2
- B1
- B2
    - Writing

Writing: rating

Rating method, rating criteria.

Source: CEFR Grid for Writing Tasks v 3.1 (ALTE)

56. Is topic known to student in advance, so that performance is rehearsed?
57. Rating method:
- holistic
- analytic: band descriptors
- analytic: checklist
- Other

58. Rating criteria (tick all that apply)
- Communicative effect
- grammatical accuracy and/or range
- lexical accuracy and/or range
- other (please specify)

59. How many raters per performance
60. Is there a procedure in case of disagreement?
61. Criteria are known to student.

Writing: task input/prompt

62. Task input/prompt

CEFR Level of Input

- A1
- A2
- B1
- B2

Control/guidance

- controlled
- semi-controlled
- open-ended

Genre of input

- letter (formal or personal)
- report
- essay
- advertisement
- other

Mode of input

- oral
- written
- visual
- a combination

63. Response elicited

Communicative purpose

- referential (telling)
- emotive (reacting)
- conative (argumentation, persuasion)
- phatic (social interaction)

Domain

- personal
- public
- occupational
- educational/academic

Register

- informal
- formal

If more than one communicative purpose, please select the main one in the drop-down menu above and add here the additional communicative purposes of the task.

Reading

Source: MUET study appendix 9

64. Reading

What is the task broadly testing?

- careful reading - local
- careful reading - global
- expeditious reading - local
- expeditious reading - global

What is the highest level of processing likely to be reached when responding to this task?

- Word recognition
- Lexical access
- Syntactic parsing
- establishing propositional meaning
- inferencing
- building a mental model,
- creating a text level representation
- creating an intertextual representation

CEFR level of reading text

- A2
- B1
- B2
- C1
- C2

CEFR domain

- personal
- public
- occupational
- educational
- other

65. Does the task contain any of these flaws?
    - assumed knowledge
    - guessable item
    - missing keys
    - unclear keys
    - incoherent item
    - other (please specify)

66. Matching Can Do statements
    - This is an open field in which you can copy and paste relevant CEFR statements which you feel this task targets.

Listening

67. Listening

What is the task broadly testing?

- careful listening - local
- careful listening - global
- expeditious listening - local
- expeditious listening - global

What is the highest level of processing likely to be reached when responding to this task?

- input decoding
- lexical search
- parsing
- establishing propositional meaning
- inferencing
- building a meaning representation
- creating a discourse representation

CEFR level of listening text

- A1
- A2
- B1
- B2
- C1

- C2
- Hard to determine

CEFR domain

- personal
- public
- occupational
- educational
- other

68. Does the task contain any of these flaws?
    - assumed knowledge
    - guessable item
    - missing keys
    - unclear keys
    - incoherent item
    - other (please specify)
69. Matching Can Do statements
    - This is an open field in which you can copy and paste relevant CEFR statements which you feel this task targets.

Structural competence

e.g. "grammar and usage", or "language in use"

70. Structural competence, item type used
    - true/false
    - multiple choice
    - gap filling (cloze)
    - ordering
    - transformation
    - sentence completion
    - short answer (word, short phrase)
    - short answer (1-3 sentences)
    - error correction (proof reading)
    - other (please specify)

Appendix 3 Team members

Project Board

Dr Nick Saville, Chair of the Project Board, Director of Research and Validation

Tim Oates, Group Director of Assessment Research and Development

Nigel Pike, Director of Assessment

Dr Ardeshir Geranpayeh, Head of Psychometrics and Data Services

Martin Robinson, Assistant Director of Assessment and Operations

Dr Hanan Khalifa, Head of Research and International Development

Stephen McKenna, Head of Communications

Core Project Team

Dr Neil Jones, Project Director

Esther Gutiérrez Eugenio, Project Coordinator

Rosey Nelson, Data Management Officer

Kasia Vazquez, Project Assistant

Dr Michael Corrigan, Analyst

Dr Joanne Venables, Analyst

Network of Experts

Hervé Marc, Regional Director Western Europe, Cambridge English Language Assessment (France)

Alistair Starling, Regional Director Northern Europe, Cambridge English Language Assessment (Germany)

Elaine Blaus, Regional Director Spain and Portugal, Cambridge English Language Assessment (Spain)

Nick Beer, Regional Director Southern Europe, Cambridge English Language Assessment (Italy)

Dr Michaela Perlmann-Balme, Referentin Entwicklung von Deutschprüfungen Bereich Sprachkurse und Prüfungen, Goethe-Institut Zentrale (Germany)

Vincent Folny, Coresponsable de la Cellule qualité et expertises au Département évaluation et certifications du Centre International d'Etudes Pedagogiques (France)

José Miguel Sánchez Llorente, Consejero delegado, Cursos Internacionales, Universidad de Salamanca (Spain)

Prof. Giuliana Grego Bolli, Professoressa di Glottologia e Linguistica e Direttore CVCL, Universita per Stranieri di Perugia (Italy)

Prof. Sabrina Machetti, Professore Associato, Dipartimento di Ateneo per la Didattica e la Ricerca, Università per Stranieri di Siena (Italy)

Content analysts

Annabelle Pinnington, Cambridge English Language Assessment (UK)

Stuart Matthews, Cambridge English Language Assessment (UK)

Bea Kálmán, Cambridge English Language Assessment (UK)

Dr Francesca Parizzi, Consultant to Cambridge English Language Assessment (UK)

Henricus Anthonius Maria Kuijper, ALTE auditor (Netherlands)

Bart Deygers, Katholieke Universiteit Leuven (Belgium)

Prof. Waldemar Martyniuk, Uniwersytetu Jagiellonskiego w Krakowie (Poland)

Dr Iwona Janowska, Uniwersytetu Jagiellonskiego w Krakowie (Poland)

Franziska Laschinger, Consultant to Goethe-Institut (Germany)

Margarete Rodi, Consultant to Goethe-Institut (Germany)

Jennifer Rosello, Consultant to Cambridge English Language Assessment (UK)

Prof. Julia Gochova Todorinova, Sofia University (Bulgaria)

Danilo Rini, CVCL, University per Stranieri di Perugia (Italy)

Claudia Buffagni, Università per Stranieri di Siena (Italy)

Marta García García, Cursos Internacionales, Universidad de Salamanca (Spain)

Dr Miriam Borham Puyal, Servicio Central de Idiomas, Universidad de Salamanca (Spain)

Comparative Judgement raters

Dr Carmela Tomé Cornejo, Cursos Internacionales, Universidad de Salamanca (Spain)

Gloria García Catalán, Cursos Internacionales, Universidad de Salamanca (Spain)

Rebeca Delgado Fernández, Cursos Internacionales, Universidad de Salamanca (Spain)

Dr Lorena Domínguez García, Cursos Internacionales, Universidad de Salamanca (Spain)

Álvaro Recio Diego, Cursos Internacionales, Universidad de Salamanca (Spain)

Alberto Buitrago, Cursos Internacionales, Universidad de Salamanca (Spain)

Marius Leonte, Cambridge English Language Assessment (UK)

Davine Sutherland, Cambridge English Language Assessment (UK)

Dr Coreen Docherty, Cambridge English Language Assessment (UK)

Victoria Watkins, Cambridge English Language Assessment (UK)
Elena Louicellier, Cambridge English Language Assessment (UK)
Cecile Loyer, Cambridge English Language Assessment (UK)
Kasia Slonka, Cambridge English Language Assessment (UK)
Jane Lloyd, ALTE Validation Unit Manager (UK)
Bea Kálmán, Cambridge English Language Assessment (UK)
David Moxon, Cambridge English Language Assessment (UK)
Wendy Bignell, Cambridge English Language Assessment (UK)
Gabriella Crawford, Cambridge English Language Assessment (UK)
Blandie Bastie, Cambridge English Language Assessment (UK)
Dr Sarah McElwee, Cambridge English Language Assessment (UK)
Diane Oliver, Cambridge English Language Assessment (UK)
Dr Kevin Cheung, Cambridge English Language Assessment (UK)
Antonio Canovas, Cambridge English Language Assessment (UK)
Gillian Horton-Krueger, Cambridge English Language Assessment (UK)
Edith Quispe Valencia, Cambridge English Language Assessment (UK)
Annabelle Pinnington, Cambridge English Language Assessment (UK)
Kate Fellows, Cambridge English Language Assessment (UK)
Louise Gilbert, Cambridge English Language Assessment (UK)
Clare Harrison, Cambridge English Language Assessment (UK)
Graham Seed, Cambridge English Language Assessment (UK)
Dr Nick Saville, Cambridge English Language Assessment (UK)
Katy Crowson, Cambridge English Language Assessment (UK)
Andrew Balch, Cambridge English Language Assessment (UK)
Andrew Kitney, Cambridge English Language Assessment (UK)
Glyn Hughes, Cambridge English Language Assessment (UK)
Lynn Stevenson, Cambridge English Language Assessment (UK)
Nick Glasson, Cambridge English Language Assessment (UK)
Crispin Davies, Cambridge English Language Assessment (UK)
Dr Nahal Khabbazbashi, Cambridge English Language Assessment (UK)
Edward Young, Cambridge English Language Assessment (UK)
Aleksandra Kledzik, Consultant to Cambridge English Language Assessment (UK)
Francesco Aiello, Consultant to Cambridge English Language Assessment (UK)
Erica Sophie Cirillo, Consultant to Cambridge English Language Assessment (UK)

Luisa Campedelli, Consultant to Cambridge English Language Assessment (UK)

Łucja Bruzda, Consultant to Cambridge English Language Assessment (UK)

Magdalena Skarbon, Consultant to Cambridge English Language Assessment (UK)

Magdalena Juras, Consultant to Cambridge English Language Assessment (UK)

Alejandro Abbud, Consultant to Cambridge English Language Assessment (UK)

Kathryn Chapman, Consultant to Cambridge English Language Assessment (UK)

Language, technical and administrative support

Graham Seed, language support

Janna Kal, language support

Simona Petrescu, language support

Jane Lloyd, technical support

Tom Gallacher, technical support

Mariangela Marulli, administrative support

Gabriella Crawford, administrative support

John Savage, proofreading

Rowan Lamb, editing and formatting

Appendix 4 Matching Can Do statements

Reading

Statement	Percentage
A1: Can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues	60%
A1: Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required	40%

Statement	Percentage
A2: Can read very short, simple texts	24%
A2: Can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus, reference lists and timetables	21%
A2: Can understand short, simple texts on familiar matters of a concrete type which consist of high frequency everyday or job-related language.	15%
A2: Can understand short simple personal letters	12%
A2: Can identify specific information in simpler written material he/she encounters such as letters, brochures and short newspaper articles describing events	9%
A2: Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items	6%
A2: Can link groups of words with simple connectors like 'and', 'but' and 'because'	3%
A2: Can understand basic types of standard routine letters, faxes and emails (enquiries, orders, letters of confirmation etc.) on familiar topics	3%
A2: Can understand everyday signs and notices: in public places, such as streets, restaurants, railway stations; in workplaces, such as directions, instructions, hazard warnings	3%
A2: Can understand regulations, for example safety, when expressed in simple language	3%
A2: Can use an idea of the overall meaning of short texts and utterances on everyday topics of a concrete type to derive the probable meaning of unknown words from the context	3%

Statement	Percentage
B1: Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension	30%
B1: Can understand texts that consist mainly of high frequency everyday or job-related language	28%
B1: Can understand the description of events, feelings and wishes in personal letters well enough to correspond regularly with a pen friend	18%
B1: Can recognise significant points in straightforward newspaper articles on familiar subjects	11%
B1: Can find and understand relevant information in everyday material, such as letters, brochures and short official documents	7%
B1: Can scan longer texts in order to locate desired information, and gather information from different parts of a text, or from different texts in order to fulfil a specific task	5%
B1: Can write straightforward connected text on topics, which are familiar, or of personal interest	2%

Statement	Percentage
B2: Can read articles and reports concerned with contemporary problems in which the writers adopt particular stances or viewpoints	49%
B2: Can understand contemporary literary prose	18%
B2: Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low-frequency idioms	10%
B2: Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms	10%
B2: Can understand specialised articles outside his/her field, provided he/she can use a dictionary occasionally to confirm his/her interpretation of terminology	6%
B2: Can obtain information, ideas and opinions from highly specialised sources within his/her field	2%

B2: Can quickly identify the content and relevance of news items, articles and reports on a wide range of professional topics, deciding whether closer study is worthwhile.	2%
B2: Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation	2%

Statement	Percentage
C1: Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of speciality, provided he/she can reread difficult sections	64%
C1: Can understand specialised articles and longer technical instructions, even when they do not relate to my field	14%
C1: I can understand long and complex factual and literary texts, appreciating distinctions of style	14%
C1: Can understand in detail a wide range of lengthy, complex texts likely to be encountered in social, professional or academic life, identifying finer points of detail including attitudes and implied as well as stated opinions	7%

Listening

Statement	Percentage
A1: Can follow speech that is very slow and carefully articulated, with long pauses for him/her to assimilate meaning	50%
A1: Can recognise familiar words and very basic phrases concerning myself, my family and immediate concrete surroundings when people speak slowly and clearly	50%

Statement	Percentage
A2: Can understand phrases and expressions related to areas of most immediate priority (e.g. very basic personal and family information, shopping, local geography, employment) provided speech is clearly and slowly articulated	32%
A2: Can catch the main point in short, clear, simple messages and announcements	20%
A2: Can understand and extract the essential information from short recorded passages dealing with predictable everyday matters that are delivered slowly and clearly	16%

A2: Can understand enough to be able to meet needs of a concrete type provided speech is clearly and slowly articulated	16%
A2: Can understand simple directions relating to how to get from X to Y, by foot or public transport	8%
A2: Can generally identify the topic of discussion around him/her that is conducted slowly and clearly	4%
A2: Can use an idea of the overall meaning of short texts and utterances on everyday topics of a concrete type to derive the probable meaning of unknown words from the context	4%

Statement	Percentage
B1: Can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure, etc.	48%
B1: Can understand straightforward factual information about common everyday or job related topics, identifying both general messages and specific details, provided speech is clearly articulated in a generally familiar accent	24%
B1: Can understand the information content of the majority of recorded or broadcast audio material on topics of personal interest delivered in clear standard speech	10%
B1: Can understand the main point of many radio or TV programmes on current affairs or topics of personal or professional interest when the delivery is relatively slow and clear	6%
B1: Can follow a lecture or talk within his/her own field, provided the subject matter is familiar and the presentation straightforward and clearly structured	4%
B1: Can follow in outline straightforward short talks on familiar topics provided these are delivered in clearly articulated standard speech	4%
B1: Can generally follow the main points of extended discussion around him/her, provided speech is clearly articulated in standard dialect	4%

Statement	Percentage
B2: Can follow extended speech and complex lines of argument provided the topic is reasonably familiar, and the direction of the talk is sign-posted by explicit markers	28%
B2: Can understand extended speech and lectures and follow even	15%

complex lines of argument provided the topic is reasonably familiar	
B2: Can understand the main ideas of propositionally and linguistically complex speech on both concrete and abstract topics delivered in a standard dialect, including technical discussions in his/her field of specialisation	15%
B2: Can understand most TV news and current affairs programmes	13%
B2: Can understand standard spoken language, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal, social, academic or vocational life	8%
B2: Can understand the majority of films in standard dialect	8%
B2: Can follow the essentials of lectures, talks and reports and other forms of academic/professional presentation which are propositionally and linguistically complex	5%
B2: Can understand most radio documentaries and most other recorded or broadcast audio material delivered in standard dialect and can identify the speaker's mood, tone etc.	5%
B2: Can understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar.	3%
B2: Can understand recordings in standard dialect likely to be encountered in social, professional or academic life and identify speaker viewpoints and attitudes as well as the information content	3%

Statement	Percentage
C1: Can follow extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly	33%
C1: Can understand complex technical information, such as operating instructions, specifications for familiar products and services	33%
C1: Can easily follow complex interactions between third parties in group discussion and debate, even on abstract, complex unfamiliar topics	17%
C1: Can understand enough to follow extended speech on abstract and complex topics beyond his/her own field, though he/she may need to confirm occasional details, especially if the accent is unfamiliar	17%

Appendix 5 Methodological notes and definitions

Additional notes on Comparative Judgement

The requirement to constrain human judgment

In a multilingual framework such as that which we have envisaged in the current project, cross-language alignment is logically seen as prior to standard setting. One would first align tests in different languages to the same scale, and only then develop interpretations – i.e. set standards. Those interpretations will then apply equally to all the aligned languages. Such considerations determined the approach to developing the Asset Languages scheme for the UK, which with 6 levels, 3 age groups and 4 skills, to be delivered for 26 languages, offered a challenge for achieving a defensible cross-language evaluation framework: human judgement should be constrained and exercised within a standardised approach.

Literature

Bramley (2005) reviews comparative approaches. The earliest of these is Thurstone's paired comparison method (Thurstone 1927), which is based on the idea that the further apart two objects are on a latent trait, the greater the probability of one of them 'winning' a comparison. Thus from a set of dichotomous judgements (e.g. of 'better' or 'worse') one can estimate not simply an ordinal ranking, but the relative location of each object on an interval latent trait scale. Thurstone's model can be implemented in different ways, of which the most computationally tractable is a Rasch formulation (Andrich 1978). However, a practical problem found by Bramley and others using paired comparisons is the repetition and sheer number of paired judgements required. A ranking approach, where more than two objects are compared, is thus an attractive alternative. One approach to this is to use rankings as categories in a Rasch partial credit model. Here the top-ranking object 'scores' 1, the second 2 and so on, for each judge involved. Bramley (2005) shows that the methods produce highly correlated results. Linacre (2006) reviews different methods of analysing rank-ordered data.

An example of a ranking approach to Comparative Judgement

The multilingual benchmarking conference organised by CIEP at Sèvres in June 2008 focused on the performance skill of Speaking. Two kinds of data were collected. At the conference itself judges rated video performances against the CEFR, with the specific feature that ratings were elicited in a 'cascade' design using English and French as 'anchor' languages: working in one group (on English and French), then in two and then three parallel subgroups, each dealing with three languages (i.e. English, French and one other).

Prior to the conference ranking data were collected from the same judges, using a web-based platform which allowed them to view video samples and record their ranking by dragging samples to re-order them in a list. The allocation of samples for the ranking exercise was such as to ensure that each judge rated in two languages, and that there was linkage in the data across all samples and languages.

Figure 46 compares the abilities estimated from rankings and ratings for the set of samples submitted to both procedures. The correlation is high. Clearly there are some significant differences in the outcomes, but given that the ranking exercise took place

before the conference, without guidance, discussion or familiarisation with the procedure, this is not surprising.

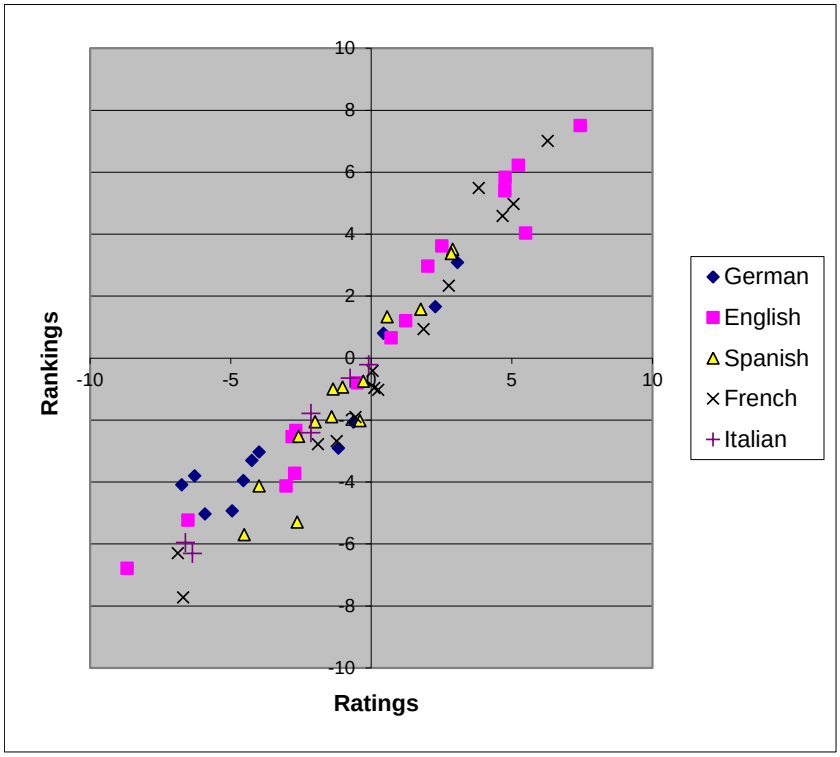
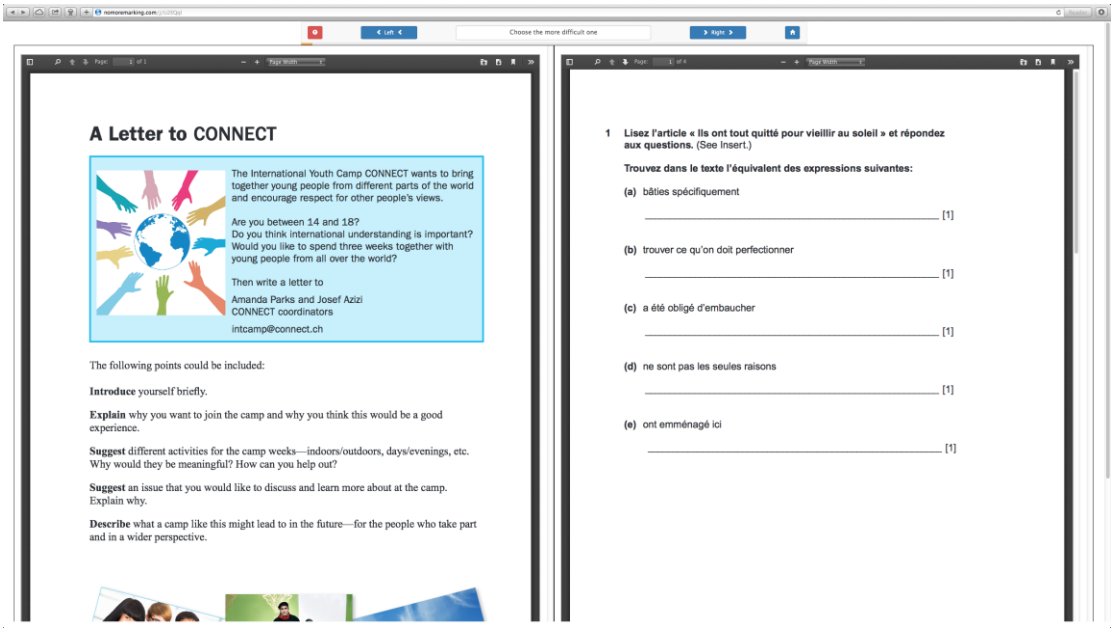


Figure 46 Ranking and rating compared (Speaking, CIEP 2008)

The *No More Marking* platform

A web-based platform called *No More Marking* was used for the CJ analysis.

The judge is presented with a series of binary choices: which of two samples is more difficult? The samples are presented side-by-side in windows which can be resized to aid readability. The judge clicks above the left-hand or right-hand window to make the choice and a new pair is instantly displayed. Judges are encouraged to make a fairly rapid and synthetic judgment.



Each judgment is recorded in a table as follows: the samples are identified by their Id numbers.

Judge	Sample Chosen	Sample Not Chosen	Time Taken
judgesname@yahoo.co.ul	12	27	00:07:23
judgesname@yahoo.co.ul	34	5	00:02:44
judgesname@yahoo.co.ul	119	59	00:01:50
judgesname@yahoo.co.ul	110	45	00:02:20
judgesname@yahoo.co.ul	26	30	00:01:18
judgesname@yahoo.co.ul	127	15	00:04:26
judgesname@yahoo.co.ul	149	115	00:02:17
judgesname@yahoo.co.ul	61	97	00:01:33
judgesname@yahoo.co.ul	98	58	00:01:30
judgesname@yahoo.co.ul	104	67	00:01:41

Analysis provided by *No More Marking*

The analysis encompasses all the samples in a given set, e.g. French Reading. For each test task it estimates a true score, which is the basis of the task's relative ranking in the dataset. The raw score is the number of comparisons which were 'won' by the task. Thus raw score plus Losses gives the number of omparisons on which the true score is estimated. The infit statistic indicates the degree of agreement between raters for each task, and thus would be of use in identifying problematic tasks.

Id	True Score	True Score SE	Infit	Raw Score	Compa-risons	Time Taken	Score	Prop Score	Losses
1	-3.72	0.4	0.87	11	138	00:32:43	11.25	0.08	127
2	-1.54	0.24	1.1	33	138	00:29:54	33.16	0.24	105
3	2.24	0.26	0.98	115	138	00:43:47	114.8	0.83	23
4	0.96	0.21	0.84	90	138	00:34:08	89.91	0.65	48
5	0.6	0.21	1.21	85	138	00:31:21	84.93	0.62	53
6	0.22	0.2	1	70	138	01:16:40	70	0.51	68
12	1.26	0.22	0.84	97	138	00:44:03	96.88	0.7	41
13	1.17	0.21	0.99	95	137	00:41:09	94.88	0.69	42
14	0.51	0.2	0.76	78	138	00:32:11	77.96	0.56	60
15	0.77	0.2	0.84	83	137	00:58:25	82.94	0.61	54
17	0.57	0.2	1.13	81	138	00:40:05	80.95	0.59	57
18	0.9	0.21	1.12	90	138	00:38:35	89.91	0.65	48
19	-0.5	0.21	1.08	55	137	00:32:42	55.06	0.4	82
20	-0.08	0.21	0.95	64	137	00:33:49	64.02	0.47	73
25	2.41	0.26	1.33	116	138	00:26:52	115.8	0.84	22

The model used: The Bradley-Terry model of Comparative Judgement

No More Marking uses the Bradley–Terry model, which gives the probability of a task winning a comparison, given the differences between two tasks. The Bradley–Terry model is a probability model that can predict the outcome of a comparison. Given a pair of items i and j drawn from some population, it estimates the probability that the pairwise comparison i > j turns out true, as

$$P(i > j) = \frac{p_i}{p_i + p_j}$$

where pi is a positive real-valued score assigned to individual i. The comparison i > j can be read as "i is preferred to j", "i ranks higher than j", or "i beats j", depending on the application.

Logit scales

The measurement scales constructed from the CJ data use units called logits. Logits scales have useful properties. In section 5 an approach was described to interpreting scores in tests by exploiting these useful properties.

Proportional scores correct are transformed into logits using the expression:

$$logit(x) = log\left(\frac{x}{1-x}\right)$$

where *x* is the proportional score.

The reciprocal expression for deriving the probable score from the logit is:

$$p = \frac{exp(a-b)}{1 + exp(a-b)}$$

where *p* is the probability of a correct response, *a* is the student's ability and *b* is the difficulty of a task. Thus probability of success relates to the *difference* between student and task on the logit scale. Where the two are identical then *p* = 50 per cent, i.e. a fifty-fifty chance. See too the introduction to Item Response Theory in Appendix 5.

Rescaling

The purpose and approach to scaling is described in 5.1.6 and 5.2.2 above.

The aproach is a linear one, which should be appropriate to the equal-interval nature of the logit scale. A common issue with the Rasch model is that scales measuring the same thing may have different means and standard deviations, which may or may not represent real differences. When combining data from different sources to construct a single measurement scale it is important to control for this.

Linear scaling uses the equation

$$Y = b(X) + a$$

where *X* are the values scaled from and *Y* are the values scaled to. *b* determines the standard deviation of the scale (the width of its units) and *a* adjusts the mean of the scale.

However, the assumption of linearity should be checked, because in practice there are issues which may impact on this. Specifically, very low or high levels of candidate performance or test task difficulty may produce stretching of the scale in the lower or higher tails. Study of the data from the Comparative Judgement, which was implemented on the No More Marking website, suggested that there was indeed a degree of stretching in the lower tail, as if particularly easy tasks were being singled out with greater probability than the Rasch model expected. The X-Y plots in sections 5.1.6 above and 5.2.2 above provide evidence of this. Thus a decision was made to remove the 3 or 4 lowest-scoring tasks from the chained anchoring procedure. This had a significant impact on the final scaling parameters used.

Figure 47Figure 19 shows the chained equating parameters for Reading and Writing, separately for English, the dual-language anchor tests, and French. Chaining works in the direction English > anchor > French. English remains unscaled, the others are scaled relative to English.

	Skill	LangGp1	A_para	B_para
1	R	Anc	-0.136	1.41
2	R Y scale		0.409	1.13
3	R X scale		0.000	1.00
4	W	Anc	-0.315	1.15
5	W Y scale		-0.549	1.01
6	W X scale		0.000	1.00

Figure 47 Scale parameters for chained equating English to French

It can be seen that some of the B parameters are quite substantial.

The second scaling, to anchor items drawn from the European Survey on Language Competences, uses parameters for English and French. In referring to the logit values

of items in the ESLC it is important to be aware that the scale for each skill and language was developed separately, and that likewise the CEFR level cutoffs were defined separately for each skill and language, as an outcome of the standard-setting conference. These issues required scaling out for the purposes of the current study.

Table 14 Parameters to scale to CEFR levels from ESLC

Parameters for English	
A_para	B_para
1.05	0.761
Parameters for French	
A_para	B_para
0.953	0.326

A simple introduction to Item Response Theory

The problem with classical statistics

The following text is taken from Jones (2014):

Figure 48 illustrates three simple statistical concepts which are familiar to anyone who has taken tests (that is, everyone): *facility*, or the mean score on a test, the *pass mark*, perhaps stated as a percentage, and the *pass rate*. Interpretation of these concepts is straightforward: as more people score more than the pass mark, so more of them will pass. But these statistics are not informative, because each of them only reflects a relationship between two underlying factors. Thus, facility reflects a relationship between the test-takers and the test: specifically, between the *ability* of the test-takers and the *difficulty* of the test items. The pass mark reflects a relationship specifically between the *difficulty* of the test items and the *standard* or passing grade which is applied. The pass rate reflects a relationship specifically between the ability of the test-takers and the standard applied.

In other words, facility, pass mark and pass rate are relative concepts with no intrinsic meaning. What we are interested in knowing is the *ability* of test-takers, the *difficulty* of the test and the level of the *standard*. These are (potentially) absolute values with intrinsic meaning: for example, a standard can be set in terms of a CEFR level, and a test-taker can be located at that level, or below or above it, by a known margin.

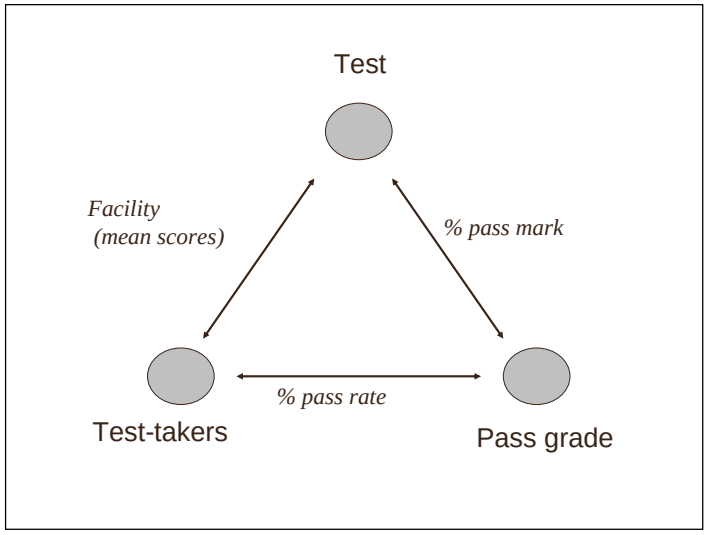


Figure 48 Three basic elements of a testing situation (after Jones & Saville, 2007)

IRT concepts

So, in an IRT view, what interests us are not scores as such, but the underlying features of learners and test items which lead to those scores being observed. The language proficiency continuum is a *latent trait* – that is, an underlying, invisible dimension - upon which learners, items and criterion levels of ability (standards) can all be located. To derive such abilities and difficulties from test response data requires the use of a specific statistical model. One widely used is the Rasch model, which belongs to a class of models within Item Response Theory.

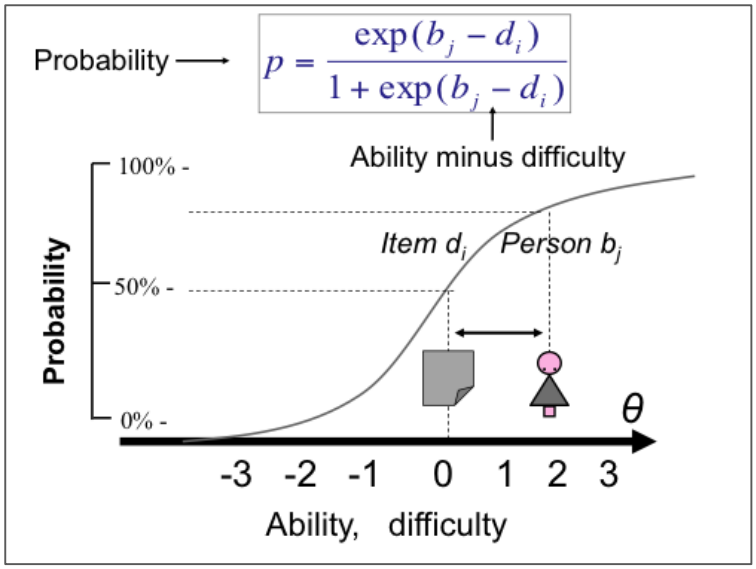


Figure 49 The Rasch model

Figure 49 includes the basic Rasch model equation, and illustrates the relationship it defines between the *probability* of a learner responding correctly to an item (the vertical y axis) and the *difference* between the item's difficulty and the learner's ability

on the horizontal axis (also called the *theta scale*, hence the Greek character on the right). Probability is a value between zero (certainly wrong) and 100% (certainly right), while the horizontal ability/difficulty scale is linear, with limits of plus or minus infinity, for total test scores of 100% or zero respectively. The scale units are called *logits*, and will be further discussed below. This explains the S-shape of the curve which describes the relationship. When a person and an item are at exactly the same point on the scale the person has a 50% chance of responding correctly. The higher the person is on the scale, the higher the probability of responding correctly (and vice versa). The relation defined by the model is quite intuitive: when the person is relatively higher on the scale than the item she is more likely than not to get it right, and when she is relatively lower she is more likely to get it wrong.

To construct a scale we must start from test data – the correct and incorrect responses given by a group of people to a group of items. The higher the total score of each person, the higher their ability. The higher the total score on each item, the lower its difficulty. That provides enough information to estimate the most likely values for all the abilities and difficulties, something that dedicated statistical software can do.

The software is necessary because estimation is not straightforward: it seeks to find the best possible fit of the data to the model, and that fit is only ever approximate. Students will get some items wrong which they would have been expected to get right, given their overall ability, and vice versa. So even after estimating the most likely ability and difficulty values for each person and item, individual responses will not be perfectly predicted by those estimated values. But this does not mean that the measurement is somehow faulty – the model works precisely because it depends on probabilities, and given enough data, probabilities can produce very accurate results. A coin is expected to land heads-up half the time, and the more throws, the closer the observed result will approach that expected outcome. Similarly, tests can produce results which are accurate to within a knowable degree of error, that error depending chiefly on the number of observations (items) in the test. Goodness of fit is important in evaluating whether the model has produced a useful measurement or not. Badly fitting data don't support substantive interpretation. Good measurement depends on well-defined constructs and well-written items, and you can only measure one thing at a time – hence the importance of testing language skills separately.

Finding the difficulty of test items is called *calibration*. Because the whole scale is defined through the relative *difference* in position between items and persons (ability minus difficulty) there is no meaningful zero point. So at the very beginning of scale construction we set an arbitrary point and ensure that every subsequent data set can be linked to it, by including some items which have already been calibrated. This is called *anchoring*. Developing suitably practical schemes for anchoring is one of the basic and most important steps in constructing a measurement scale.

The above description shows that in an Item Response Theory view, ability and difficulty define each other: they arise in the interactions of learners and tasks. This notion is in fact clearly analogous to a socio-cognitive view of validity, where ability is seen to reflect observable interactions between the cognition and skills of a learner and the demands of a task. Good model fit may thus strengthen the claim for the *interactional authenticity* of test tasks.

In thinking about measurement scales it is worth trying to keep separate in our minds the measure, which is a number indicating a point on a proficiency scale, and the thing measured, which reflects cognitive attributes of the learner, as elicited by content attributes of the tasks. Of course, our focus in testing is on the learners, but the test tasks define completely what we can expect to discover about them.

So the term "proficiency" is defined here in terms of a measure, and interpretations drawn on the basis of the measure. Proficiency thus defined does not exist until someone measures it. Thus we must distinguish it from terms identifying various kinds of ability or competence and so on which may be used in defining the construct of what is tested. These describe posited properties of learners which exist independently of whether they are measured or not.

The argument for the validity of measures eventually comes back to our theoretical model of cognition and the interactions with test tasks that we predict we will observe, given the features designed into them. To the extent that test performance empirically confirms these predictions then our claim for the validity of the test is strengthened.

How task difficulty, ability and standards relate

As explained above, a scale co-locates three things: test tasks, learners, and standards, that is, points on the scale which indicate achievement of some criterion-referenced level. Taking the levels of the CEFR as examples of standards we can see how these three notions interact.

The tasks define what is tested. Grouping tasks by level allows us to characterise each level in terms of the sort of things learners can do. But learners define levels too: we understand levels not only in terms of *what* things learners can do but also *how well* they can do them. Another way of looking at Figure 49 above defines a point on a proficiency scale in terms of task difficulty *plus* performance level.

Performance level is more easily understood in relation to the performance skills of Writing or Speaking: a task such as *describing your holiday* does not relate strongly to a level. It sounds like an appropriate kind of task for a learner at CEFR A2 or B1, but every level of performance on the task is imaginable. Performance on item-based tests such as of Reading and Listening is more simply evaluated – items are answered right or wrong. Here performance level must be understood in terms of the probability of getting an item at a certain level correct. The expected total score in a test is simply the sum of the probabilities across items, so that higher performance will relate to higher total scores.

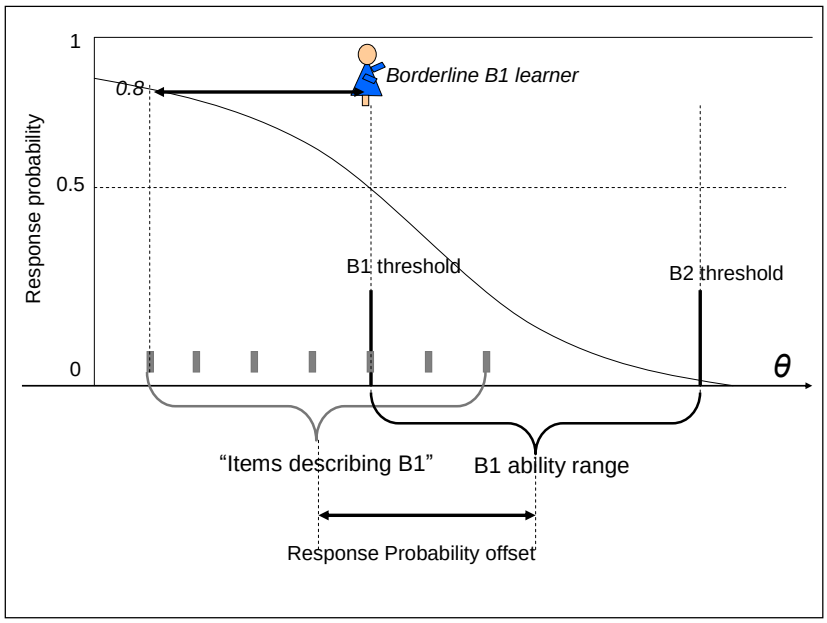


Figure 50 Items, learners and levels on a measurement scale

Figure 50 illustrates two levels: CEFR B1 and B2. Each level has a lower threshold at which it is achieved, and a higher threshold where the next level takes over. A borderline B1 learner is shown. This learner has just achieved B1 level, and will continue to be at B1 until she achieves the next level up.

What do we expect of a B1 learner, even a borderline one, in terms of being able to do the kind of tasks which describe B1? We recognise that there are easier and harder B1 tasks, because B1 covers the whole range of difficulty between the end of A2 and the beginning of B2. We would expect the borderline learner to have mastered the easiest B1 tasks but not the harder ones. But thinking of mastery creates an apparent problem, because as illustrated above, this learner has only a 50% chance of responding correctly to an item at the same point on the scale as she is, and her chance on more difficult items is even lower. This does not square with our idea of mastery – surely she should have a much higher chance (i.e. *response probability*) on the easiest B1 items? An 80% probability is a frequent rule-of-thumb definition of mastery, although this is of course an arbitrary choice.

The conclusion is clear: the tasks which we take to describe B1 level reflect an expectation that learners at that level will be able to perform them reasonably well. This is true both of objective test tasks and performance-based tasks such as Writing. An adequate performance level is built into our understanding of the task. Thus the B1 level threshold defines the point at which learners can be said to be "at B1 level", but it is confusing to talk of *items* as being "at a level". Better to speak of describing the level, or providing information about learners at the level. In terms of their location on the measurement scale items which we take to describe the level will be offset downwards from the level thresholds. The size of that offset reflects a judgement about the response probability which we choose to specify as indicating minimum mastery.

Item banking

Item banking is a methodology for constructing tests and interpreting test outcomes using an IRT model. Its great value is that it creates an interpretive framework that encompasses exams at different levels, over different exam administrations and test versions, making it possible to generate tests with very similar measurement characteristics and to grade them to constant standards. Figure 51 gives a schematic view of item banking as a methodology for test construction.

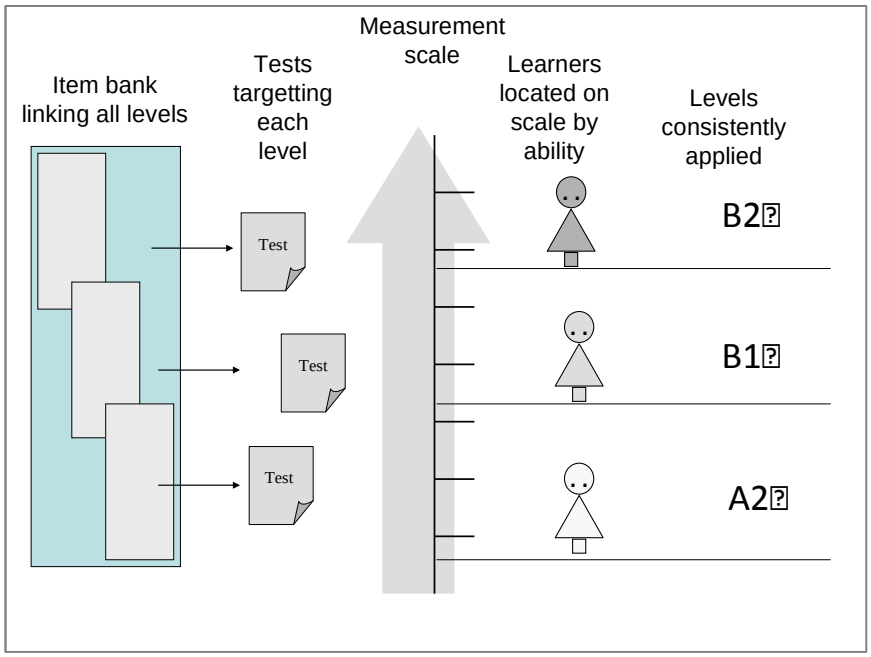


Figure 51 Item banking approach to scale construction and use (after Jones & Saville, 2007)

Figure 51 shows on the left an item bank containing tasks ready for use in a test. The difficulty of the items in each task is known, that is, they have been calibrated. The data to calibrate these tasks has come from some form of pretesting, and we calibrated them to a single scale by using anchor tests, administered to pretest candidates together with the pretests themselves.

With the item bank stocked, tests are assembled by selecting tasks of appropriate difficulty for the target levels. Candidates' scores on tests locate them on the measurement scale according to their ability. Figure 51 shows tests at three levels, and three candidates. Although they might all have the same score – say, 70%, we know that 70% on the easiest test indicates a lower ability than 70% on the hardest test: knowing the item difficulties enables us to locate the candidates precisely on the measurement scale.

Finally, the standards are applied. These are fixed points on the scale which can be directly applied to establish each candidate's grade. Even if test versions differ slightly in difficulty, the standard can be held constant. If we modify the standard, then that will impact all future tests in the same way.

Figure 51 thus illustrates the power of a fully-functional item banking system. In such a system *ad hoc* standard setting is neither necessary nor possible. The great benefit of an item banking approach is not simply that it facilitates the construction of a stable measurement scale, but that in consequence it facilitates the construction of *meanings* which explain what it is that the scale measures.

Firstly, the items in the bank provide a concrete, detailed description of progression in terms of test content. Secondly, the fact that standards can be precisely maintained from session to session and from level to level facilitates doing the research to develop stable interpretations of learners' performance in the world beyond the test – for example in Can Do statements such as those used in the descriptive scales of the CEFR.

Thirdly, standards may described in linguistic terms. The English Profile (http://www.englishprofile.org) is a large-scale study which has produced a linguistic description of CEFR levels, identifying salient features of each level based on an extensive corpus of learner performance data (Hawkins & Filipovic, 2012). All such developments exploit and contribute to the meanings embodied in the measurement scale.

Interpreting national/regional levels of performance

In the body of this report we have illustrated the possibility of linking the difficulty of a particular jurisdiction's test tasks to CEFR levels. This is a necessary step on the way to interpreting the performance of students in tests: knowing the difficulty of tasks on a scale linked to the CEFR enables us to interpret the performance of students, also in terms of CEFR levels.

Most jurisdictions were not able to provide summary performance data in a form which could readily be exploited for this purpose. However, we can illustrate on the example of one (anonymous) country, for whose ISCED 3 level test data are available as explained below.

The tables in section 0 above show that it is possible to determine the overall difficulty of a country's test tasks on a CEFR-related scale. A jurisdiction would need to be able to estimate the mean difficulty of all the test tasks for a given test, on the CEFR-linked scale (the scale emerging from the CJ exercise can be seen as a first attempt at such a scale, and a CJ procedure might be a good way for jurisdictions to use such a scale to anchor the level of an exam).

Knowing the mean difficulty of the test would then allow scores on the test to be interpreted, and converted to CEFR-linked grade levels.

Table 15 Illustration of linking grades to CEFR scale

1. Number achieving each grade	2. Grades	3. Percentage achieving	4. As cumulative fraction	5. Logit
906	1	6.6	0.07	-2.65
962	2	7	0.14	-1.85
1169	3	8.5	0.22	-1.26

1297	4	9.5	0.32	-0.77
1381	5	10.1	0.42	-0.34
1587	6	11.6	0.53	0.13
1533	7	11.2	0.65	0.60
1497	8	10.9	0.75	1.12
1282	9	9.3	0.85	1.71
905	10	6.6	0.91	2.35
769	11	5.6	0.97	3.44
407	12	3	1.00	6.91

Let us say that this country's test has a mean difficulty level of 2.5. This is the logit value where a student at the same point on the scale would have a likelihood of scoring 50 percent. Students of higher ability will score above 50 percent, while those of lower ability will score below 50 percent, in scale units which reflect the relationship of ability to the probability of achieving a particular score (the logit scale). In this way we may attempt to model the ability of the groups achieving each of the 12 grades in the test.

Table 15 Illustration of linking grades to CEFR scale shows how approximate logit values can be derived from score data. Columns 1 and 2 comprise the data provided by the country. It shows twelve grades, and the number of students achieving each grade. To these we have added:

- the percentage achieving each grade or higher (which is derived from column 1);
- the same thing rendered as a cumulative fraction;
- the cumulative fraction transformed into a logit (see Appendix 5 for how this is done).

Appendix 6 The Common European Framework of Reference for Language

The following table is taken from the text of the CEFR. It provides a brief description of the CEFR levels A1 to C2.

Table 1. Common Reference Levels: global scale

Proficient User	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices
Independent User	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.

	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

NC-01-15-650-EN-N

Publications Office

ISBN: 978-92-79-50995-7