Study on comparability of language testing in Europe

Executive summary

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

Executive summary

September 2015

Directorate-General for Education and Culture

2015 ERASMUS+ **EN**

This document has been prepared for the European Commission; however, it reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

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 978-92-79-50997-1

doi: 10.2766/987189

Printed in Belgium

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, as shown in Table 1 of the report (section 2.2.1).

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. 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 8.2 below.

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 that results are similarly valid and reliable across all jurisdictions.**

NC-04-15-633-EN-N

ISBN: 978-92-79-50997-1