Engagement in different language testing and assessment activities: - Were there differences between regions of Europe?

ENLTA / EALTA survey, Part 2

Detailed findings regarding the respondents’ current and past engagement in different assessment activities are reported below, activity by activity. Only those respondents who worked in Europe and replied to the questions in Part 1 of the questionnaire (Teachers’ Questionnaire) are included here (n = 542 – 582, depending on the question / activity).

Classroom-focused testing / assessment

Preparing your own classroom tests

Most of the respondents to Part 1 of the Questionnaire, 70-85%, were engaged in this activity at the time of taking part in the survey, and almost all of the rest had at some time been engaged in this activity.
No statistically significant differences between the regions were found.

Using ready-made tests from textbook packages or from other sources
Engagement in this activity varied from region to region: most of the respondents were or had been engaged in this but often only half of them were currently using ready-made tests. However, in some regions 90% of the respondents to the teachers’ questionnaire used such tests. Very significant differences were found between the regions (chi square = 88.6, df = 12, p = .000). In particular, the Eastern European and Baltic respondents were much more often engaged in this activity than could be expected, and conversely, the Northern and Western Europeans much less often.

Interpreting test results (classroom or external tests)

Not surprisingly, almost everybody was currently engaged in this activity, i.e. 80-90% of the respondents. No statistically significant differences between the regions were found.
Giving feedback to students based on information from tests / assessment

As with the previous activity (interpreting test results), almost all respondents to Part 1 were engaged in this activity. No statistically significant differences between the regions were found.

Using self / peer assessment
Self and peer assessment was quite common: 60-70% of the respondents were currently engaged in it and further 10% had at some time used it. However, at least 20% had never used it.
No statistically significant differences between the regions were found.

Using informal, continuous, non-test type of assessment

[Graph showing engagement levels by region]

Current engagement in this activity varied between 60 and 80%, depending on the region, and further 10-20% had used it previously.
Significant differences were found between the regions (chi square = 21.8, df = 12, p = .040). In particular, the Eastern European respondents chose the option ‘engaged only in the past’ less often than could be expected, and conversely, the Baltic respondents chose it more often than could be expected.

Using the European Language Portfolio, an adaptation of it or some other portfolio
Portfolio turned out to be the least used of the classroom-oriented assessment activities: only 20–40% of the respondents had ever used it. This also turned out to be the type of assessment where, overall, the need for further training was the highest. Significant differences were found between the regions (chi square = 22.5, df = 12, p = .033). In particular, the Northern European respondents chose the option ‘engaged only in the past’ more often than could be expected. Also, many of the differences in the frequency of choosing the ‘engaged now’ option seemed to contribute to the overall significance of the group differences.

**Purposes of testing**

To give grades
Perhaps not surprisingly, almost everybody were or had been engaged in giving grades to learners, and 80-90% of the respondents to Part 1 of the Questionnaire were doing that at the time of replying to the survey questions. Significant differences were found between the regions (chi square = 22.5, df = 12, p = .033). In particular, the Eastern European respondents chose the option ‘engaged only in the past’ less often than could be expected, and conversely, the Southern Europeans more often than could be expected.

To find out what needs to be taught / learned next

![Bar chart showing engagement in teaching and learning needs across regions of Europe](image)

Almost as many of the respondents as above (grade giving) were engaged in formative / diagnostic assessment: 70-80% were engaged in it currently. No statistically significant differences between the regions were found.

To place students onto courses, programmes, etc
Compared with grade giving and formative assessment above, placing students appeared to be a purpose of assessment in which ‘only’ half of the respondents were currently engaged in – further 10-20% had at some point done that, however. Very significant differences were found between the regions (chi square = 57.7, df = 12, p = .000). In particular, the Southern and Eastern European respondents were engaged in this activity significantly more often than could be expected, and conversely, the Northern and South-Eastern Europeans less often than could be expected.

To award final certificates (from school / programme; local, regional or national level)

As in the case of placing students, this activity was also somewhat less common: about 60% of the respondents were engaged in this. However, very significant differences were found between the regions (chi square = 47.9, df = 12, p = .000). In particular, the South-
Eastern European respondents were engaged in this activity significantly less often than could be expected, and conversely, the Southern, Eastern and Western Europeans more often than could be expected.

Content and concepts

Receptive skills (reading/ listening)

Not surprisingly, practically everybody was or had been engaged in assessing reading or listening. No statistically significant differences between the regions were found.

Productive skills (speaking/writing)
As in the case of receptive skills above, the great majority of the respondents were engaged in assessing speaking and/or writing. No statistically significant differences between the regions were found.

Microlinguistic aspects (eg grammar/vocabulary)

[Bar chart showing engagement levels across regions of Europe]

Somewhat similarly to the above macro-skills, most of the respondents were assessing microlinguistic skills, too. However, significant differences were found between the regions (chi square = 47.9, df = 12, p = .000). In particular, the Western European respondents chose the option ‘engaged only in the past’ more often than could be expected, and conversely, the Eastern Europeans less often than could be expected.

Integrated language skills
Assessment of language skills in an integrated fashion turned out to be a very common activity, too, with some 80% of the respondents being engaged in it currently. The proportion of those who had never done it was somewhat higher than in the case of assessing the skills separately above. (Whether that is statistically significant or not has not however been studied.) No statistically significant differences between the regions were found.

Aspects of culture

Of the types of content in language testing/assessment, culture seemed to be the least frequent: only some 50-60% of the respondents were assessing it. However, very significant differences were found between the regions (chi square = 47.9, df = 12, p = .000). In particular, the Northern Europeans were more often engaged in this activity than
could be expected, and the Southern Europeans less often. Also, the Western European respondents chose the option ‘engaged only in the past’ more often than could be expected, and conversely, the South-Eastern Europeans less often.

Establishing reliability of tests / assessment

![Chart showing engagement levels in establishing reliability across regions of Europe]

About half of the respondents to the teachers’ version of the questionnaire were currently engaged in establishing the reliability of their assessments, and further 10-20% had done that previously. Since using statistics is very central in establishing reliability of assessments, it is interesting to compare the responses to this activity with those given to ‘using statistics’ below. We can notice that the proportion of respondents who reported being engaged in using statistics is somewhat lower (30-40%) than the proportion who are engaged in establishing the reliability of assessments. It is likely that the respondents have interpreted the notion of ‘establishing reliability’ as being more general than just studying e.g. inter-rater reliability or the homogeneity of tests with the help of statistical analyses. Possibly, many respondents included conscious efforts to ensure reliability in their assessments into this activity although they may not have used statistics as part of the process. Another non-statistical activity that could be seen part of ‘establishing reliability’ is taking part in such training where common standards and interpretations are sought.

No statistically significant differences between the regions were found.

Establishing validity of tests / assessment
Very similarly to the question on reliability, about 50% of the respondents reported being currently engaged in some activity that aimed at establishing the validity of their assessments. On the basis of the results of this survey it is impossible to elaborate on how different respondents may have interpreted ‘validity’. Since the concept and meaning of validity has been the subject of a lot of discussion and debate in the professional literature, it is quite probably that different interpretations exist among language teachers as to what validity means and what counts as ‘establishing validity’. In fact, the last point might make an interesting topic for further research by members of EALTA.

No statistically significant differences between the regions were found.

Using statistics to study the quality of tests / assessment
Only a minority, some 30-40% of the respondents reported being currently engaged in using statistics to study the quality of tests and assessments. Typically, over 50% had never used statistics (at least, not for the purpose at hand). See above for a brief discussion on the relationship of this activity to ‘establish reliability’. No statistically significant differences between the regions were found.

External tests and examinations (regional or national)

Taking part in rating oral or written performances

Of the activities related to external tests and examinations, taking part in the rating of the students’ performances was one of those in which quite a lot of the respondents were either currently or previously engaged in (50-60% were currently engaged in this activity). The other equally common activity was being an interviewer or interlocutor in such tests, which, in fact, often combines interviewing and rating. No statistically significant differences between the regions were found.

Using statistics to study the quality of tests / assessment
Only a minority (20-30%) of the respondents were engaged in this activity. No statistically significant differences between the regions were found.

Writing items/test tasks for an examination body

Item writing for external examination was an activity in which 25-35% of the respondents were engaged in currently, and further 10% in the past. No statistically significant differences between the regions were found.

Reviewing items/test tasks for an external examination
Reviewing items for external exams was almost as rare as writing them: some 25-40% were engaged in it.

No statistically significant differences between the regions were found. (However, the overall significance almost reached the .05 level (p = .062, chi square = 20.2, df = 12). In particular, the Western and Southern European respondents seemed to differ from the others.)

Acting as an interviewer / interlocutor in an oral test or examination

Together with rating speaking or writing, this was the activity related to external exams that most of the respondents were or had been engaged in.

Significant differences were found between the regions (chi square = 33.0, df = 12, p = .001). In particular, the South-Eastern Europeans chose the option ‘not at all engaged’ more often than could be expected and the Central Europeans less often. Also, the
Western and Central European respondents chose the option ‘engaged only in the past’ more often than could be expected.

Defining assessment criteria

Some 35-50% of the respondents were engaged in ‘defining assessment criteria’ in the context of external tests and examinations. This is one of the activities which may, in hindsight, be rather difficult to interpret. The fact that so many of the respondents reported engagement in this activity makes one wonder how this was interpreted by the respondents. Quite often, presumably, external test and examination boards define their assessment criteria and scales (within a small circle of employees) and then issue them to be used by e.g. the teachers or employees of the board who administer the tests. More information on the kind of interpretations given by the respondents to this activity would be useful before any firm conclusions could be drawn about the meaning of the responses given to this particular activity.

No statistically significant differences between the regions were found.