RELANG

Relating language examinations to the common European reference levels of language proficiency: promoting quality assurance in education and facilitating mobility

Marking, Rating and Grading

European Centre for Modern Languages and European Commission cooperation on INNOVATIVE METHODOLOGIES AND ASSESSMENT IN LANGUAGE LEARNING

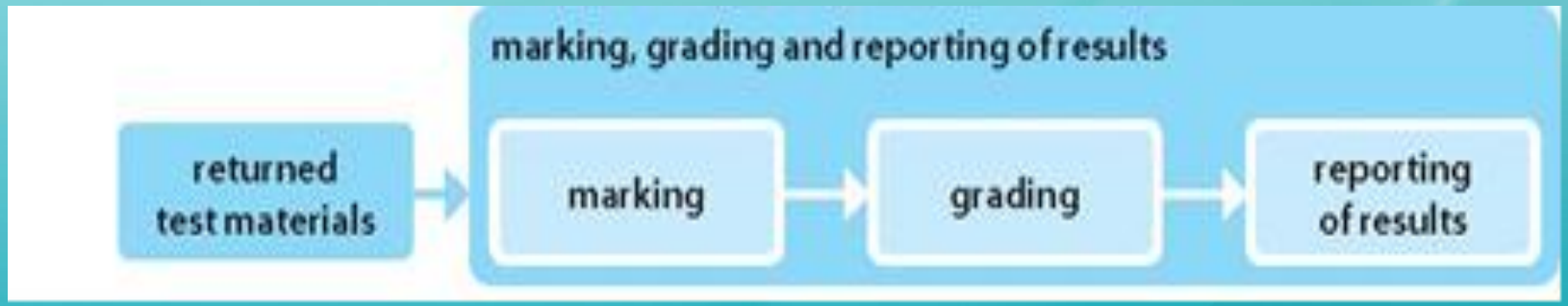
EUROPEAN UNION

EUROPEAN CENTRE FOR MODERN LANGUAGES
CENTRE EUROPÉEN POUR LES LANGUES VIVANTES

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

The Process

The three steps at this stage are represented below:



Marking

- While the expression *marking* covers all activities by which marks are assigned to test responses, a distinction is often made between the *marker*, indicating a less skilled role, and the *rater*, which is a role requiring professional training.
- We also distinguish *clerical* (i.e. human) and *machine* marking.

Rating

- This is marking where the exercise of trained judgement is necessary, to a much greater degree than in clerical marking.
- When judgement is used, a single ‘correct answer’ cannot be clearly prescribed by the exam provider before rating. For this reason, there is more scope for disagreement between judgements than in other kinds of marking, and thus a greater danger of inconsistency, between raters, or in the work of an individual rater.

Rating Scales

- This is a set of descriptors which describe performances at different levels, showing which mark or grade each performance level should receive.
- Rating scales reduce the variation inherent in the subjectivity of human judgements.

Types of Rating Scales (1)

- **Holistic or analytic scales:** a single mark for a performance can be given using a single scale describing each level of performance.
- **Relative or absolute scales**: scales may be worded in relative, evaluative terms (e.g. ‘poor’, ‘adequate’, ‘good’), or may aim to define performance levels in positive, definite terms.
- **Checklists:** marks based on a list of yes/no judgements as to whether a performance fulfils specific requirements or not.

Types of Rating Scales (2)

- **Generic vs. task-specific scales:** An exam may use a generic scale or set of scales for all tasks, or provide rating criteria which are specific to each task. A combination of both is also possible.
- **Comparative vs. absolute judgement:** It is possible to define a scale through exemplar performances; the rater's task is to say whether a performance is lower, higher or the same in relation to one or more exemplars. A mark is thus a ranking on a scale, e.g. in terms of CEFR levels.

Rating Process

- Raters must have a shared understanding of the standard. The basis of this shared understanding is shared examples of performance.
- For small-scale exams a group of raters may arrive at a shared understanding through free and equal discussion.
- For large-scale exams the standard must be stable and meaningful: experienced examiners with authority to communicate the standard to newcomers.

Rater Training

Training should proceed through a series of steps from more open discussion towards independent rating, where the samples used relate to the exam being marked:

- guided discussion of a sample, through which markers come to understand the level
- independent marking of a sample followed by comparison with the pre-assigned mark and full discussion of reasons for discrepancies
- independent marking of several samples to show how close markers are to the pre-assigned marks.

Grading

- In language tests that report results in terms of CEFR levels, grading needs to be *criterion-referenced:* performance is evaluated with respect to some fixed, absolute criterion or standard.
- An exam may be designed to report over several CEFR levels, or just one. In the latter case, those test takers who achieve the level may be said to have ‘passed’, and the others to have ‘failed.
- Identifying the score which corresponds to achieving a certain level is called *standard setting*.