A QUALITY MATRIX FOR CEFR USE: Examples of promising practices

1 OVERVIEW

Project leader(s) contact: Armin Berger

Country: Austria

Institution: University of Vienna

Type of context: National

Educational sector: Tertiary

Main focus: Test

SUMMARY

Name: Academic writing and speaking – ELTT initiative

Abstract:
This document describes the Austrian University English Language Testing and Teaching (ELTT) initiative, developed by the Language Testing Centre at Klagenfurt University, with the aim of professionalising assessment practices for high-stakes examinations in Austrian university English language programmes. The initiative involved construct definition and scale development/validation projects for the purposes of certifying academic writing and speaking proficiency at the end of BA programmes.

Stage: Evaluation

Theme: Assessment

CEFR aspects used: Levels, descriptors, assessment with defined criteria

Main features of this example:

- Cyclical rating scale development based on CEFR descriptors, a committee's expertise and experience, identification of key concepts in sample performances, iterative rating of performances and refinement of descriptors, involvement from external experts, and benchmarking
- Thorough, documented development process, using the CEFR
- Systematic validation employing a mix of methods, including descriptor sorting, descriptor calibration, and descriptor-performance matching
- Thorough statistical validation (classical test theory and multi-faceted Rasch analysis)
- Collaboration with different stakeholders (e.g. university language teachers, linguists and students)
- Focus on the impact of assessment

Quality principles particularly demonstrated: Validity, Transparency, Coherence

European Centre for Modern Languages of the Council of Europe
www.ecml.at/CEFRqualitymatrix

2 PROJECT DESCRIPTION

Background:
This project arose from a perceived need to standardise learning/teaching objectives and outcomes in higher education, to increase transparency, to be able to show and increase coherence in the language programme, to facilitate cooperation among teachers, to be able to communicate aims and assessment results more effectively.

Stated aims:
The main aim of the project was to professionalise assessment practices for high-stakes examinations in Austrian university English language programmes. The focus was on the certification of writing and speaking proficiency at the end of the B.A. programmes. There was also a strong interest in the impact of assessment practices on teaching and learning in university language programmes. The intended outcome was a set of analytic rating scales at levels C1 and C2 for the assessment of academic writing, presentations and interactions, plus benchmarked performances for rater training as well as teaching and learning purposes.

Steps/stages:

Rating scale development:

a. *Familiarisation*: Prior to the first scale development workshop, the participants were sent a document with CEFR descriptors for familiarisation purposes. They were asked to study the descriptors and refresh their knowledge of the communicative activities and competences at levels C1 and C2. The actual workshop started with a brief consultation session in which the participants compared the courses and language testing practices in Austrian English departments, deepened their understanding of the CEFR's specifications at C1 and C2, and received some expert input on different procedures for developing rating scales.

b. *Establishing the design principles*: The initial workshops were devoted to establishing the design principles. While some of these principles had been decided in advance, including the analytic approach to scoring, the number of dimensions per rating scale and the design methodology, others were up for discussion. It was decided that each scale should comprise four criteria subdivided into six levels, with the top band being based on C2 descriptors and the bare pass on C1 descriptors:

   - Band 1 (top) Defined with C2 descriptors
   - Band 2 undefined
   - Band 3 Defined (see below: step d)
   - Band 4 undefined
   - Band 5 Defined with C1 descriptors
   - Band 6 undefined – but below C1

   In addition to the rating scales, the team aimed to produce benchmarks for academic writing and speaking for each scale criterion and for as many bands as possible, accompanied by written justifications for the scores awarded.

c. *Specification*: After the design principles had been established, the group set out to define the test construct. In a brainstorming activity, first in small groups and then in a plenary session, the participants identified relevant assessment criteria. For the presentation scale, the criteria include lexico-grammatical resources and fluency, pronunciation and vocal impact, structure and content, and genre-specific presentation skills. For the interaction scale, the criteria are lexico-grammatical resources and fluency, pronunciation and vocal impact, content and relevance, and interaction skills.

   The next step was to establish subcategories of each criterion. To this end, the category labels were

written as headings on posters, and small teams were asked to write down key areas for each dimension. The results were reported back to the plenary, discussed and decided upon. Once the criteria and their key features had been identified, they were transformed into draft descriptors for the top and the bare pass level. The team extracted suitable anchor descriptors at C1 and C2 from the CEFR's Tables 1-3 and scales of illustrative descriptors. Where these descriptors proved insufficient or inadequate, the team extended or adjusted them. New descriptors were written to cover those areas that were felt to be undertreated in the CEFR descriptors.

d. *Componential analysis*: The primary purpose of this stage was to refine the draft descriptors and define band 3 mid-way between C1 and C2. The method adopted was qualitative in nature, involving the identification of key concepts based on sample performances; the team selected typical performances at different levels, identified key features of each performance and incorporated them into the descriptors. The product of this stage was a set of analytic rating scales with three defined bands from C1 to C2.

e. *Operational analysis and benchmarking*: In the final workshops, the focus shifted from descriptor formulation to trialling and benchmarking. The aim of this stage was to identify prototypical performances for as many scale criteria and bands as possible, while at the same time confirming the soundness and applicability of the scale descriptors. In an on-site rating session, the team rated and discussed a number of sample performances. First, all team members rated the performances on all criteria and wrote down brief justifications for their ratings. In a subsequent plenary session, the participants compared and discussed their interpretations. Then the raters were invited to reconsider their initial decisions in the light of the discussion. A performance was considered benchmarked if a consensus of n-1 was reached. In parallel, a few minor adjustments to the descriptor formulations were made, mainly to hone the wording and sharpen the boundaries between the bands. Small local teams finalised the written justifications for the benchmark scores.

Rating scale validation (speaking scales)

a. *Descriptor sorting*: In a separate project, the speaking scales (*academic presentations* and *interactions*) were validated in three stages. In the first stage, a descriptor sorting task was conducted. The rating scale descriptors were divided into independent, minimally meaningful descriptor units. A number of qualified university teachers of English, who had not been involved in the development process, were asked to sort the descriptor units into different bands of proficiency. Correlation analyses including both consistency and agreement indices were conducted to determine the strength of the relationship between the judgements.

b. *Descriptor calibration*: As a next step, the sorting task data was subjected to a multi-faceted Rasch analysis[1] to take account of different facets of the test situation, including, most notably, rater variability.

c. *Descriptor-performance matching*: Then the rating scale descriptors were linked to samples of real speech. Expert teachers were invited to indicate the extent to which individual descriptor units represented a given student performance. The data collection was organised in three steps.

[1] Multi-faceted Rasch measurement (MFRM) is a variant of the Rasch model (which is often used to construct item banks for tests). Instead of just two 'facets' (item, candidate) for which difficulty and ability (respectively) are estimated, MFRM defines a third facet – the assessor, measures their severity/leniency and takes it into account when estimating the ability of the candidate. It is thus a way of 'objectifying' subjective assessment. MFRM is usually operationalised in the program FACETS, developed by Mike Linacre, its founder. Further facets can also be defined. MFRM was used in the Swiss research project that developed the CEFR descriptors and in the various benchmarking seminars that developed DVDs of spoken CEFR performance samples. A good introduction (from the reference supplement to the Council of Europe's Manual for relating tests and examinations to the CEFR) is provided here: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a23

- First, an outcome-based role-play discussion task was trialled with undergraduate students to ensure that it elicited the expected responses.
- Second, video recordings of student presentations and discussions were produced in a series of mock exams.
- Finally, experienced university language teachers were asked to match the descriptor units with the video performances using a questionnaire. Again, multi-faceted Rasch analysis was used to analyse the data.

d. *Rating scale revision*: The findings from the three previous stages were synthesised to reintegrate the most effective descriptor units into improved versions of the scales. A systematic evaluation procedure was established to create a quality hierarchy of descriptor units, based on the soundness of the calibrations, statistical model fit, consistency across procedures in terms of band allocation after setting cut-off points at equal intervals, and congruence with the scale developers' original band allocation. This process resulted in two revised rating scales with calibrated descriptors at five bands from C1 to C2.

Operational trial (speaking scales)

a. *Quantitative analysis*: Multi-faceted Rasch analysis was used to confirm the functionality of the revised scales. Experienced language teachers rated a number of videorecorded performances on all dimensions of the scales under realistic assessment conditions. The focus of the analysis was on student separation, rater severity and consistency, criterion difficulty as well as rating scale effectiveness.

b. *Qualitative analysis*: Two retrospective group interviews with the raters were conducted after the rating session in order to investigate their perceptions of how the revised scales function under operational conditions.

Timeline:
October 2006 – July 2008: Development of writing scale
October 2008 – July 2010: Development of speaking scales
2012-2014: Validation and revision of speaking scales
2017: Operational trial of speaking scales
Ongoing: Stronger linkage between our language programme and the CEFR is desired, but has not fully materialised yet due to resource constraints.

People/roles:
University language teachers, linguists and students from the Universities of Graz, Innsbruck, Klagenfurt, Salzburg and Vienna; directors of studies were addressed; collaboration throughout the process, with a strong and enthusiastic coordinating team.

Other resources needed:
Material resources: rooms for regular scale development workshops, rooms and projectors at five Austrian English Departments for mock exams with students, video recording equipment.
Additional human, time and financial resources would be needed for the validation and refinement of the writing scale, the full implementation of the scales, systematic rater training and the examination of washback effects.

Publications that have been used or produced related to this example:
Berger, Armin; Heaney, Helen, (forthcoming). "Developing rating instruments for the assessment of academic writing and speaking at Austrian English departments". In Sigott, Günther; Cesnik, Hermann (eds.). *Language testing in Austria: Taking stock. / Sprachtesten in Österreich: Eine Bestandsaufnahme*. Frankfurt am Main:

Peter Lang.

Berger, Armin (forthcoming). “Rating scale validation for the assessment of spoken English at tertiary level”. In Sigott, Günther; Cesnik, Hermann (eds.). *Language testing in Austria: Taking stock. / Sprachtesten in Österreich: Eine Bestandsaufnahme*. Frankfurt am Main: Peter Lang.

Berger, Armin (2016). “Rating scales for assessing academic speaking: A data-based definition of progression“. *VIEWS* 25, 25-44.

Berger, Armin (2015). *Validating analytic rating scales: A multi-method approach to scaling descriptors for assessing academic speaking*. Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang.

3 RESULTS

What was achieved: A set of analytic rating scales for *writing, academic presentations* and *interactions* as well as a set of benchmarked performances.

Impact:
The project received positive reactions from project participants, teachers and international experts and was praised as an unprecedented step in professionalising assessment practices in Austrian university language departments. At the same time, there is concern about the limitations of the approach and its practicality for local contexts; systematic implementation is still lacking. Colleagues started to revise other existing rating instruments and procedures, but more systematic dissemination and implementation of the project results would be desirable. Currently, the initiative has lost momentum.

Resources on this theme:
http://www.uni-klu.ac.at/ltc/inhalt/430.htm

4 ADVICE AND LESSONS LEARNT:

- Follow procedures aimed at systematic implementation. Build a good and highly motivated team.

- Don’t just focus on the development of specific instruments, but also on their implementation and the “bigger picture”. Consider the time after the project, next steps, future developments etc. carefully.

- Watch out for dependence on volunteers in place of allocated resources**:** The validation phase depended on the idealism of one researcher and a number of raters who spent hours rating performances for nothing in return. However, there were no resources for the implementation phase.