Publié dans Bulletin VALS-ASLA, n° spécial, tome 3, 233-240, 2015,
source qui doit être utilisée pour toute référence à ce travail

The CEFR, language norms and testing

Curtis GAUTSCHI
University of Bern
Center for the Study of Language and Society
Länggassstrasse 49, 3000 Bern, Switzerland

ZHAW Zurich University of Applied Sciences
School of Applied Linguistics
Theaterstrasse 17, 8401 Winterthur, Switzerland
gaut@zhaw.ch

Der Einfluss des Gemeinsamen Europäischen Referenzrahmens für Sprachen (GER/CEFR) auf den Fremdsprachenunterricht in Europa und darüber hinaus ist unbestritten. Der GER wirkt sich grossflächig auf die Entwicklung und Koordination von Lehrbüchern, Programmen, Lehrplänen und der Ausbildung von Lehrpersonen aus sowie in besonderem Masse im Prüfungswesen im Bereich Fremdsprachen (Dose et al. 2010: 2). Es ist darum von wesentlicher Bedeutung, die Angemessenheit dieser Wirkung mit Berücksichtigung der deklarierten Ziele und Eigenschaften des Referenzrahmens kontinuierlich zu überprüfen. Seit seiner Veröffentlichung wurde in verschiedener Hinsicht Kritik am GER geübt; insbesondere wurden in einigen Bereichen seine Unvollständigkeit bemängelt und Präzisierungen gefordert (Weir 2005), während andere Arbeiten potentielle sprachpolitische Missbräuche ansprachen (Fulcher 2004). Nur wenig hinterfragt wurde hingegen bis jetzt die grundsätzliche Eignung des GER für das Testen und die Zertifizierung von Sprachkenntnissen, vor allem im *high-stakes General Purposes testing*, einer seiner Hauptanwendungen. In diesem theoretischen Beitrag wird diskutiert, wie der GER angemessen verwendet werden kann und inwiefern er als Instrument für allgemeine Sprachprüfungen, von denen für die Lernenden viel abhängt, geeignet ist. Besonders berücksichtigt werden dabei die ursprünglichen Zielsetzungen und die strukturellen Grundeigenschaften des GER.

Stichwörter:
Gemeinsamer Europäischer Referenzrahmen, GER, General Purposes Testing, Sprachkenntnisse, Prüfungsvalidität, Real World Assessment, Validitätstheorie.

1. Does the CEFR set norms?

Before examining issues related to the proper use of the CEFR, it is necessary, in the interest of fairness, to discuss briefly what the CEFR claims and does not claim. While there is no doubt that the CEFR has had a massive impact, "rapidly becoming a powerful instrument for shaping language education policies in Europe and beyond" (Martyniuk 2011: 34), the CEFR also makes broad and explicit disclaimers regarding the setting of norms in practice. As Martyniuk (2010: viii) states, the text itself, as well as other documents issued by the Council of Europe and other related language policy projects, "heavily downplay(s) the notion that the CEFR offers standards." CEFR readers are constantly reminded that "Users of the Framework may wish to consider and where appropriate state:" areas that are of particular importance to users. Thus, in answer to the question *Does the CEFR set norms?*, its authors and mandating bodies have taken pains to respond with a clear "no" – "The CEFR

is purely descriptive – not prescriptive, nor normative." (Council of Europe 2008: 9).

However, if one asks a slightly different question, *Does the CEFR define norms?*, the answer is different, insofar as descriptions of levels of language proficiency constitute norms if they are operationalized. The CEFR (Council of Europe: 2001:1) makes this clear through the strong claim that it "describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively".

Upon reflection, it may become obvious that this is indeed the ultimate objective – by comprehensively defining everything a learner or teacher needs, the CEFR effectively defines the foreign language proficiency construct for learning and testing. The authors of the CEFR are aware that this is desirable, noting that it "is already clear however that a set of common reference levels as a calibrating instrument is particularly welcomed by practitioners ... who ... find it advantageous to work with stable, accepted standards of measurement and format." (Council of Europe 2001: xiii). The attempt to define what practitioners are already aware of, but of what they perhaps lack a systematic expression, was a design purpose in the research projects that led to the CEFR's development as North (2000: 1) states: "The Swiss research project sought to make transparent teachers' "fuzzy" internalised norms and standards".

The foreign language teaching and testing field in the European context has responded with near unanimous acceptance of the CEFR's description of the foreign language proficiency construct, as evidenced by its broad use in the development and linking of textbooks, syllabi and curricula, teacher education and training, as well as foreign language ability assessment (Dose et al. 2010: 2). Thus, we may conclude that, while the CEFR itself clearly does not force its use or application, educational institutions and foreign language practitioners themselves have operationalized the CEFR, and in doing so, have set norms and standards for teaching in testing.

2. Is the CEFR appropriate for all types of FL testing?

With regards to testing in particular, there is an ever-increasing demand for accurate language assessment on the part of employers, government institutions and other interested parties. The FL testing industry, especially general proficiency or general purposes testing, has naturally expanded to meet these demands. As mentioned above, the strong claim on the part of the CEFR to "provide descriptors of communicative language proficiency" (North 2000: 2) has led to the CEFR's current position as the dominant force in testing, "dictating the construct in assessment projects throughout Europe" (McNamara 2003: 471). This is evidenced, specifically in the Swiss context,

through the linking of foreign language assessment to CEFR bands. This can be found at the public school system level, including mandatory secondary level schools and tertiary-level institutions, such as professional schools, universities and universities of applied sciences (CDIP 2011, Consortium HarmoS Langue étrangères 2009, SUPSI 2012), at the adult continuing education level (e.g. Eurocentres, Migros Club Schools and other private institutions), at the governmental level (e.g., FIDE Federal Office for Migration framework for the linguistic integration of migrants - see Lenz et al. 2009), and the private language testing industry in general (e.g., Cambridge English, IELTS, Goethe Institut examinations, Diplôme d'études en langue française, TELC).

In light of its broad use, the question we wish to raise here is *Is the CEFR appropriate for all the different types of FL testing it is used for*?

2.1 Two types of testing: education system testing and real-world gateway testing

To answer the question regarding the appropriateness of the CEFR for different types of language testing, we must first distinguish between two contexts where test outcomes have meaning or value. Some tests, for instance, are intended for use within the classroom or broadly speaking, the education system. Examples include placement tests, end of course tests, tests that bridge the transition from one school level to another, or from one educational setting to another, from one country to another, and so on. In these cases, after the test, the test taker finds him-/herself either in the same classroom as before or in another. Since test outcomes are used and interpreted within the education system, we may call this *education system testing*.

In contrast, after some tests, test takers do not find themselves in a classroom, but in the real world, or, put differently, the language ability tested will be used in contexts outside of classroom settings. Here, test outcomes are intended to have meaning for stakeholders outside the classroom, for example, employers or immigration boards (see FIDE framework - Lenz et al. 2009). Here, tests are intended to measure language ability for contexts of use that are beyond the classroom. To distinguish this type of testing from *education system testing*, let us call it *real-world gateway testing,* since this type of test intends to measure language ability for use beyond the classroom and is intended to have predictive validity with respect to real-world performances (Davies et al. 1999: 149). We may thus refine our question: *Is the CEFR appropriate for both education system testing and real-world gateway testing*?

3. Validation argument

The question of appropriateness of an assessment instrument can only be properly answered by examining the validation argument in favour of its use. One may counter that validation evidence is required for a test, not a framework, and that the final responsibility for validation rests solely on the test developer rather than the CEFR, if it is used. However, since a framework can have a) *a test development function* (since it can be used to *define* the proficiency construct in a test directly), b) *an external criterion function* (since it can be used to *check* the proficiency construct of a test), and c) *a comparability function* (which allows it to *compare* the proficiency construct of two tests) – all of which the CEFR is intended for (Coucil of Europe 2001: 1-20, 182) – validation evidence is most certainly required.

Since a testing instrument's validity is based on the evidence that correct test outcome interpretation is possible in its context of use, with respect to its a) design purpose or intended use, and its b) design properties ‑ including data used, decisions made, informants and methods (Messick 1989: 13, Hughes 1989: 26, American Educational Research Association 1999: 9, Bachman & Palmer 1996: 22, Fulcher 2003: 117), let us examine the CEFR's design purpose and design properties.

3.1 CEFR design purposes

The CEFR's design purposes are clearly documented and are summarized here. The text states that the framework was designed to (Council of Europe 2001: 1-20):

(1) facilitate comparisons between different educational stages or systems regarding
   a) learning objectives
   b) courses syllabi and levels
   c) materials and materials development
   d) tests and examinations
   e) achievement and qualifications
   f) curriculum guidelines

(2) enhance the comparability (transparency) of
   a) points (1) a) to f) above

(3) promote and facilitate co-operation among educational institutions in different countries

AND

(4) provide explicit description of learning objectives and learning content at
   a) class level
   b) course level
   c) program level
   d) curriculum level
   e) institutional level
   f) inter-institutional level

(5) impact teaching methods

(6) allow learners' achievement/progress to be measured at each stage of learning along continuum of learning

(7) facilitate planning of self-directed learning

(8) enable transportability of language qualifications to different educational contexts aid educational

(9) assist system inspectors

FOR TESTING:

(10) define content syllabus of tests and examinations

(11) define assessment criteria "both in relation to the assessment of a particular spoken or written performance, and in relation to continuous teacher-, peer- or self- assessment."

(12) describe the levels of proficiency "in existing tests and examinations thus enabling comparisons to be made across different systems of qualifications."

As can be seen from this summary, the CEFR's design purposes are strongly linked to the classroom/learning/education system context.

3.2 CEFR design properties

With regards to the CEFR's design properties, we may consult North (2000), who describes the research project that led to the development of the proficiency scales that represent the core of the CEFR.

Briefly, the project consisted of three phases. In the initial intuitive phase, over 1,600 proficiency descriptors from prior existing rating scales were collected, analyzed and categorized. All of the source scales, with the single exception of Fulcher's (1996) empirically derived fluency descriptors, were all developed using language-expert intuition rather than any empirical study of contexts of language use.

In the qualitative phase, teacher participants' discussions of learners' video performances were analyzed to verify the categories and descriptors. This was followed by 32 workshops with teacher participants to a) sort descriptors into analytic categories b) judge their usefulness to teachers (clarity, relevance) and c) sort them into proficiency bands.

In the quantitative phase, teachers assessed learners using questionnaires containing the descriptors derived from the qualitative phase. A year later, teachers assessed learners again, this time with different questionnaires. Using multi-faceted Rasch scaling, the common scale of proficiency descriptors was created. These scales are now found in chapters 3, 4 and 5 of the CEFR. The result is a comprehensive framework that defines the various aspects of language ability across all language levels, and provides a common language for educational practitioners to use to locate language learners along the continuum of language learning.

Summarizing the design process, we note, however, that final descriptors are basically a synthesis of prior existing rating scales primarily based on language-expert intuition, teachers were the only informants, teacher utility was the main developmental principle, and the identification of "anchor points" along the proficiency continuum were all defined based on teacher evaluations of learner performances.

3.3 Call for research

In conclusion, we may say with confidence that the CEFR's stated purposes and design characteristics support the validation argument for its use in teacher/learner education system contexts. Its use for other purposes, however, in particular for *real-world gateway testing* as described here, lacks rigorous empirical support and requires external verification based on a study of the perceptions of proficiency of those representative of the context of post-test language use. This is the focus of a current three-year (2014-2017) joint study of the University of Bern and ZHAW. In this study, we compare the communicative language ability construct related to speaking ability, as perceived by a broad range of participants representative of the various test score use domains implicit in General Purposes Speaking testing to that of language professionals. Based on the qualitative data collected through an analysis of verbal protocols recorded in reaction to pre-recorded speaking test performances, assessment criteria relevant to the different domain groups of interest will be identified. These criteria will subsequently be used in a quantitative phase, via an assessment questionnaire, to compare how various sets of criteria regress on assessments of overall communicative language ability in the different groups of interest. The main outcome of this multi-method study will be the evaluation of the validation argument in favour of current assessment criteria.

As a general call for research, it is suggested that avenues such as the following be explored for the purpose of improving the meaningfulness of test scores:

- The development of a *Context-of-Language-Use* approach to validation, whereby validation arguments are evaluated based on the perceptions of proficiency of interlocutors within the context where post-test language will be used
- research using a real-world approach and real-world informants to empirically define the proficiency construct
- CEFR adaptation: calibration using real-world informants

This is all in line with the CEFR's own invitation to modification, verification and validation of its use. The CEFR states that it is "open: capable of further extension and refinement" and "dynamic: in continuous evolution in response to experience in its use" (Council of Europe 2001: 8). Further, the "CEFR is context-neutral – it needs to be applied and interpreted with regard to each specific educational context in accordance with the needs and priorities specific to that context" (Council of Europe 2008: 9). Finally, while the CEFR attempts to be comprehensive, it cannot, of course, claim to be exhaustive. Further elaboration and development are welcomed (Council of Europe 2008: 9).

4. Conclusion

In this paper, we have briefly discussed the FL learning norm-defining properties of the CEFR and evaluated the CEFR's appropriateness for use in both *education system testing* and *real-world gateway testing*, with regards to its design purposes and design properties. We conclude that, while there is design validation evidence for its use in education system settings, this is lacking for non-classroom settings and therefore further research is required to support its use for such purposes, including general proficiency testing. In the interest of good practice and the desire to provide meaningful information regarding test takers' abilities to all stakeholders, we urge further research in the direction of a context-of-language-use approach to language performance test validation.

REFERENCES

American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington DC.: American Educational Research Association.

Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: OUP.

Conférence suisse des directeurs cantonaux de l'instruction publique (CDIP). (2011). *Compétences fondamentales pour les langues étrangères*. Retrieved from http://edudoc.ch/record/96779/files/grundkomp_fremdsprachen_f.pdf

Consortium HarmoS Langue étrangères. (2009). *Langues étrangères. Rapport scientifique de synthèse et modèle de compétences.* Retrieved from http://www.edudoc.ch/static/web/arbeiten/harmos/L2_wissB_25_1_10_f.pdf

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge*:* Cambridge University Press.

Council of Europe. (2008). *Recommendation CM/Rec(2008)7 of the Committee of Ministers to member states on the use of the Council of Europe's Common European Framework*

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & Mcnamara, T. (1999). *Dictionary of Language Testing. Cambridge University Press.*

Dose, S., Götz, S., Brato, T. & Brand, C. (2010). *Normen in Educational Linguistics*. Frankfurt/M: Peter Lang GmbH, Europäischer Verlag der Wissenschaften.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13* (2), 208-238.

Fulcher, G. (2003). *Testing Second Language Speaking*. London: Longman.

Fulcher, G. (2004). Deluded by Artifices? The Common European Framework and Harmonization. *Language Assessment Quarterly: An International Journal*, *1*(4), 253-266.

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Lenz, P., Andrey, S. & Lindt-Bangerter, B. (2009). *Rahmencurriculum für die sprachliche Förderung von Migrantinnen und Migranten. University of Fribourg Institute of Multiculturalism.* Retrieved from http://www.institute-multilingualism.ch/assets/files/ipl/Rahmencurriculum-d-201003 def.pdf

Martyniuk, W. (2010). *Aligning Tests with the CEFR: Reflections on Using the Council of Europe's Draft Manual.* Cambridge: Cambridge University Press.

Martyniuk, W. (2011). The Council of Europe's Common European Framework of Reference for Languages (CEFR): Approach, status, function and use. *Language Learning in Higher Education*, *1*(1), 23-40.

Mcnamara, T. (2003). Looking back, looking forward: Rethinking Bachman. *Language Testing*, 20, 466-473.

Messick, S. (1989). Validity. In R. Linn (ed.), *Educational Measurement*. New York: Macmillan/American Council on Education, 13-103.

North, B. (2000). *The Development of a Common Framework Scale of Language Proficiency*. New York: Peter Lang.

SUPSI. (2012). *Direttive interne del Centre for Languages and International Relations.* Retrieved from http://www.supsi.ch/clir/il-centro/direttive.html

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