Language Value 2014, Volume 6, Number 1 pp. 15-27
http://www.e-revistes.uji.es/languagevalue Copyright © 2014, ISSN 1989-7103

Issues in aligning assessments with the Common European Framework of Reference

Spiros Papageorgiou
spapageorgiou@ets.or
Educational Testing Service, MI, USA

ABSTRACT

One of the main aims of the Common European Framework of Reference is to help providers and users of assessments "describe the levels of proficiency required by existing standards, tests and examinations in order to facilitate comparisons between different systems of qualifications" (Council of Europe 2001: 21). Providers of language assessments both inside and outside Europe follow various methodologies to align their assessments with the CEFR levels, as several case studies show (Figueras and Noijons 2009; Martyniuk 2010). This paper discusses the use of the CEFR in the field of language assessment, focusing in particular on issues related to alignment. The paper presents the types of validity evidence that should be accumulated to support an alignment claim and concludes with directions for future research in order to further enhance our understanding of using the CEFR for the design of assessments and the interpretation of assessment results.

Keywords: *Alignment, cut scores, Common European Framework of Reference.*

I. INTRODUCTION

The publication of the Common European Framework of Reference (CEFR) in 2001 has been recognized as the "most significant recent event on the language education scene in Europe" (Alderson 2005b: 275). The main purpose of the CEFR is to provide a common basis for the elaboration of language syllabuses, examinations, and textbooks by describing in a comprehensive way what language learners have to learn to do in order to use a language effectively for communication (Council of Europe 2001: 1). The language proficiency levels and their language performance descriptors are central to the CEFR's descriptive scheme of language use, as noted by Little (2006: 169). They serve one of the main aims of the Council of Europe as described in Chapter 3 of the CEFR volume, that is: "to help partners to describe the levels of proficiency required by existing standards, tests and examinations in order to facilitate comparisons between

DOI: http://dx.doi.org/10.6035/LanguageV.2014.6.3

different systems of qualifications" (Council of Europe 2001: 21). Such comparability of language qualifications in Europe was difficult to achieve prior to the CEFR because of the plethora of diverse educational systems and traditions. Alderson (2007: 660) pointed out that "the six main levels of the CEFR have become a common currency in language education, and curricula, syllabuses, textbooks, teacher training courses, not only examinations, claim to be related to the CEFR".

Nowadays, providers of language assessments, both inside and outside Europe, follow various methodologies to align their assessments with the CEFR levels, as reported in several case studies in two edited volumes (Figueras and Noijons 2009; Martyniuk 2010). The most common approach to bring tests into alignment with the CEFR is the one recommended in the *Manual* published by the Council of Europe (2009). The approach consists of two main stages: content alignment and setting of cut scores. The main purpose of this paper is to discuss the use of the CEFR in the field of language assessment, with a particular focus on issues related to the alignment of assessments with the CEFR. Before discussing alignment issues, however, it is important to first consider the work that led to the development of the CEFR and its levels, which is presented in the next section.

II. THE DEVELOPMENT OF THE CEFR AND ITS LANGUAGE PROFICIENCY SCALES

The Council of Europe (not be confused with the European Union) is the continent's oldest political organization, founded in 1949. Its general aim is to foster common democratic principles among its 47 members. The Council of Europe has been active in the area of languages for more than forty years with two complementary bodies: the Language Policy Division in Strasbourg, France, and the European Centre for Modern languages in Graz, Austria.

In order to promote plurilingualism and pluriculturalism among European citizens, the Council of Europe published a number of documents in the 1970s that have been influential in second language teaching, such as the notional-functional syllabus by Wilkins, which describes what a learner communicates through language (Wilkins 1976), and three ascending levels describing language achievement: Waystage (Van Ek

and Trim 1991), Threshold (Van Ek and Trim 1998) and Vantage (Van Ek and Trim 2001). In 1991, at an intergovernmental symposium in Switzerland, the development of a common framework for learning, teaching and assessment was deemed desirable in order to:

- promote and facilitate cooperation among educational institutions in different countries;
- provide a sound basis for the mutual recognition of language qualifications;
- assist learners, teachers, course designers, examining bodies and educational administrators in situating and coordinating their efforts;

(Council of Europe 2001: 5)

The authoring group produced an initial version in 1996 and the final version of the CEFR was published after feedback and consultation in 2001, the European Year of Languages, in English and French. Since then, the CEFR volume has been freely available online on the Council of Europe website (www.coe.int/portfolio) in more than 30 languages. These include non-European languages, such as Arabic and Japanese, revealing the strong interest in the document world-wide, not only within Europe.

Although the CEFR contains a rich description of the language learning process, it is widely accepted that the CEFR language proficiency scales are the best known part of the 2001 volume (Little 2006). The proficiency scales of the CEFR have gained popularity because they offer a comprehensive description of the objectives that learners can expect to achieve at different levels of language proficiency. They describe language activities and competences at six main levels: A1 (the lowest) through A2, B1, B2, C1 to C2 (the highest). Borderline levels are further elaborated using a 'plus' between A2+ (between A2 and B1), B1+ (between B1 and B2) and B2+ (between B2 and C1). The scales comprise statements called 'descriptors', which are always phrased positively, as they are intended to motivate learners by describing what they can do when they use the language, rather than what they cannot do (Council of Europe 2001: 205). The performance descriptors of the CEFR are designed following an action-oriented approach: language users are seen as members of a society who have tasks to accomplish, including those that are not language-related (Council of Europe 2001: 9).

Because of the action-oriented emphasis, the descriptors are also frequently referred to as "can-do statements".

The scales and descriptors in the 2001 edition of the CEFR were primarily developed during a large research project in Switzerland (North 2000; North and Schneider 1998). The project applied a variety of qualitative and quantitative methods for the initial analysis and collection of more than 2000 language descriptors used in proficiency scales around the world, the consequent selection and refinement of 1000 of these descriptors and, finally, the placement of the descriptors at different proficiency levels that subsequently formed the CEFR levels (see also Appendix A and Appendix B in Council of Europe, 2001). A number of studies and research projects such as the DIALANG project (Alderson 2005a; Alderson and Huhta 2005) have shown that the descriptors can be consistently replicated in a range of contexts, thus offering validity evidence for the use of those descriptors across a variety of contexts.

Language assessment is specifically discussed in Chapter 9 of the CEFR, which serves as a useful introduction to important notions and principles in the field. Fundamental terms such as validity and reliability are explained, and different types of assessment are described (e.g., formative versus summative assessment; norm-reference testing versus criterion-referencing testing). The next section focuses on the process of aligning assessments with the CEFR, which has been the topic of many studies in the field of language assessment.

III. THE PROCESS OF ALIGNING ASSESSMENTS WITH THE CEFR

The CEFR has been extremely influential in the field of language assessment, as evidenced by the 2005 special issue of the Language Testing journal on language assessment in Europe (Alderson 2005b) and the various alignment studies in the two edited volumes mentioned above (Figueras and Noijons 2009; Martyniuk 2010). The demand for alignment of assessments with various external standards has increased not only in Europe, but worldwide, because of education reforms which push for accountability, including close monitoring of students' progress and use of standardized tests (Deville and Chalhoub-Deville 2011). In the United States, the No Child Left Behind Act and more recently the Common Core State Standards, an initiative

supported by most states in the United States to describe the skills and abilities expected by students at each grade level, have further raised the demand to bring assessments into alignment with frameworks and standards.

The *Manual* published by the Council of Europe (2009) offers a recommended set of procedures for aligning tests with the CEFR, which consists of two main stages: content alignment and setting of cut scores. For content alignment, the Specification chapter of the *Manual* suggests forms to be completed for each language skill. These forms contain several questions regarding the extent to which the content of an assessment covers communicative language activities, contexts, text types and other aspects of language ability described in the CEFR. Thus, the completed forms constitute a claim of content coverage in relation to the CEFR. The second stage involves the setting of minimum scores on the test that would indicate that a test-taker has demonstrated the performance expected at that CEFR level (Standardization Training and Benchmarking chapter and Standard Setting Procedures chapter). These minimum scores (cut scores) are established following a well-researched process in the educational measurement literature called "standard setting" (Cizek and Bunch 2007). During standard setting, a panel of expert judges (often called "panelists") is required, under the guidance of one or more meeting facilitators, to make judgments on which examination providers will base their final cut score decisions. Statistical information about test items and the distribution of scores might also be used to help panelists with their judgment task. A fairly common practice in standard setting meetings is that more than one round of judgments is implemented (Hambleton 2001; Plake 2008). Between rounds, the panel discusses individual judgments, receives the statistical information about items and scores and repeats the judgments. Even though the panel will offer a recommended cut score, the decision on whether to accept this score rests with the examination provider. In this sense, standard setting is in fact a procedure for recommending cut scores for implementation by the provider of the test (Cizek and Bunch 2007; Tannenbaum and Katz 2013). Procedures for validating the recommended cut scores are also presented in the *Manual*.

IV. EXPLORING THE QUALITY OF ALIGNMENT WITH THE CEFR

In the field of educational measurement, alignment typically refers to the extent to which the content of an assessment covers the skills and abilities described in an external framework and standards. Such exploration of content coverage is an integral part of the Specification chapter in the *Manual*. Webb (2007) proposed a process to evaluate the alignment of assessments with content standards based on four criteria:

- Categorical Occurrence, which addresses the issue of whether a test covers the content discussed in the standard.
- Depth-of-Knowledge (DOK) Consistency, which addresses the extent to which an assessment is as cognitively challenging for test-takers as one would expect, given the description of what students are expected to know and be able to do in the standard.
- Range of Knowledge Correspondence, which deals with the extent to which the breadth of knowledge in the assessment corresponds to the expected one in the standard.
- Balance of Representation, which addresses the extent to which specific knowledge is given more or less emphasis in the assessment compared to the standard.

Although the *Manual* (Council of Europe 2009) does not provide criteria similar to the ones by Webb (Webb 2007) for evaluating alignment of test content with the CEFR, it could be argued that the various forms that should be completed during the Specification stage do cover the above criteria to some extent.

Content alignment as described in both the *Manual* (Council of Europe 2009) and Webb (2007) requires the use of human judgment. This use of human judgment is a central issue in the process of setting cut scores (Zieky and Perie 2006: 7). As Kantarcioglu and Papageorgiou (2011) noted, judgments are not only involved during the planning of a standard setting meeting, for example, when a standard setting method is chosen, but in every step of the decision-making process of setting cut scores, that is: deciding on the number of levels with which to classify test-takers, selecting and training panelists, and scheduling the activities in the standard setting meeting. Despite this reliance on

judgments, the standard setting meeting and its outcomes can be evaluated based on several criteria typically grouped into three categories (Council of Europe 2001; Hambleton and Pitoniak 2006; Kane 1994):

- Procedural validity, examining whether the procedures followed were practical, implemented properly, that feedback given to the judges was effective, and that sufficient documentation has been compiled.
- Internal validity, addressing issues of accuracy and consistency of the standard setting results.
- External validation, by collecting evidence from independent sources which support the outcome of the standard setting meeting.

The *Manual* presents in detail how (mostly) quantitative data under these three categories should be collected and analyzed to support the proposed cut scores in relation to the CEFR levels. Studies reporting on the alignment of assessments with the CEFR routinely employ such quantitative techniques to provide validity evidence for the setting of cut scores in relation to the CEFR levels. For example, the alignment of the reading and listening scores of the Michigan English Test with the CEFR levels (Papageorgiou 2010b) involved examination of both intrajudge and interjudge consistency, such as standard error of judgment, agreement coefficient, and Kappa indices as part of the internal validation of the cut score. In another study, Kantarcioglu et al. (2010) applied the many-facet Rasch model (Linacre 1994) to explore the judges' agreement in setting cut scores for the Certificate of Proficiency in English of Bilkent University to the CEFR levels.

A qualitative approach to investigating the judges' decision-making process when setting cut scores to the CEFR was employed by Papageorgiou (2010a). The study investigated the factors reportedly affecting the panelists' decision to set a cut score and the problems faced when setting cut scores in relation to the CEFR. The panelists' group discussions were analyzed based on a coding scheme built both inductively, that is, drawing codes from the actual data, and deductively, that is, drawing codes from existing theory, such as qualitative research into participants' experiences in standard setting (Buckendahl 2005). The findings of the study suggest that decision-making might be affected by factors irrelevant to the description of expected performance in the

CEFR, such as panelists' personal expectations and experiences, which might threaten the validity of the cut score. The study also found that the CEFR might be useful for defining learning objectives, but is not sufficiently specified for the purpose of setting cut scores.

To conclude, research approaches that evaluate alignment with the CEFR include both quantitative and qualitative techniques and there are a growing number of relevant studies employing both. However, future research still needs to address several issues regarding CEFR alignment, as discussed in the next section.

V. FUTURE RESEARCH IN CEFR ALIGNMENT

When aligning test scores with the CEFR, an important decision to be made is whether a score that demonstrates sufficient performance on the assessment also indicates sufficient performance in relation to the CEFR. This is particularly the case for language assessments reporting results in the form of a pass/fail result. A pass/fail result is usually the case with licensure examinations, intended for professionals such as doctors or pilots, who need to pass the exam in order to practice their profession. Language assessments might also report pass/fail results, typically accompanied by a certificate which documents that a test-taker performed satisfactorily on the assessment. If the content of this assessment has been aligned with a specific level on the CEFR, then the implication is that all test-takers with a "pass" certificate should be at the intended CEFR level. Therefore, two decisions need to be made regarding the use of the scores from such an assessment: first, whether a score indicates that a test-taker has passed the assessment, and second, whether this "pass" score indicates that the targeted CEFR level has been achieved (see Council of Europe 2009: 58). More research is needed to understand the relationship between these two cut score decisions, which for now remains unclear.

Aligning assessments with the CEFR has important implications for policy-making. There has been considerable criticism of the uses of the CEFR as a policy document (McNamara 2006; McNamara and Roever 2006), in particular when it comes to immigration. According to Alderson (2007: 260), an unintended consequence of the adoption of the CEFR as a tool by policy-makers is that these officials have no

understanding of the nature of language learning, yet they impose requirements for language proficiency without any consideration as to whether such levels are achievable. For example, language tests are extensively used as gatekeepers for immigration purposes (Shohamy and McNamara 2009) based on language requirements defined in terms of CEFR level (see for example requirements by the UK Border Agency at http://www.ukba.homeoffice.gov.uk). However, the rationale behind the selection of a given CEFR level for a specific purpose such as immigration is not always clear. Therefore, more research is needed in local contexts to identify reasonable language requirements for specific language uses in order to inform policy-making.

Another important implication of CEFR alignment for learners, teachers, and score users is the interpretation of results from different assessments that claim alignment with the same CEFR level. These different assessments should not be interpreted as equivalent in terms of difficulty or content coverage (Council of Europe 2009: 90). Achieving CEFR Level B1 on a general proficiency test intended for young learners and a test of English for Specific Purposes (ESP) does not mean that the scores on these two tests have the same meaning because the intended test purpose, test content, and test-taking population are notably different. One way to provide more accurate information about assessment results is to provide empirically-derived, test-specific performance levels and descriptors designed for a given assessment, for example by adopting a scale anchoring methodology (Garcia Gomez, Noah, Schedl, Wright, and Yolkut 2007). Such levels and descriptors can be provided in addition to information about CEFR alignment.

VI. CONCLUSION

As discussed in this paper, the CEFR and in particular its language proficiency scales and descriptors might offer language teachers, learners and users of assessments an opportunity to better understand the meaning of the results of these assessments (Kane 2012). However, alignment with the CEFR should not be considered a substitute to ongoing procedures for validation (Fulcher 2004). The *Manual* strongly emphasizes that a prerequisite for any effort to achieve alignment with the CEFR is that an assessment be of high quality, otherwise alignment is "a wasted enterprise" (Council of Europe

2009: 90). For example, if an assessment is not reliable, setting a minimum score on this assessment to indicate adequate performance at a given CEFR level will not be particularly meaningful. Moreover, it should also be pointed out that the theoretical underpinnings of the CEFR remain weak (Alderson 2007) and that its language proficiency scales are primarily a taxonomy that makes sense to practitioners, rather than empirically validated descriptions of the language learning process (North and Schneider 1998: 242-243). Moreover researchers have noted several problems with the use of the CEFR for designing test specifications (Alderson et al. 2006). Therefore, content alignment of an assessment with the CEFR cannot provide sufficient evidence of content validity or substitute various language learning theories that should be considered when designing an assessment.

Alignment with the CEFR might not be straightforward because, by design, the description of what learners are expected to do in the CEFR is under-specified to allow for a wider application across a variety of contexts. Unfortunately, this intended under-specification might also mean that alignment of assessments for specific groups of test-takers may be particularly challenging, for example, in the case of assessments for young learners (Hasselgreen 2005). Despite these issues, it could be argued that alignment of assessments with the CEFR remains an important area of inquiry in the field of language assessment because it has the potential to raise awareness of important assessment issues, for example in contexts where local tests are developed (Kantarcioglu, et al. 2010).

Acknowledgements

The author would like to thank his ETS colleagues Don Powers, Veronika Timpe, and Jonathan Schmidgall for their useful comments on an earlier version of this paper. The author is responsible for any errors in this publication. Any opinions expressed in this paper are those of the authors and not necessarily of Educational Testing Service.

REFERENCES

Alderson, J. C. 2005a. *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C. 2005b. Editorial. *Language Testing, 22*(3), 257-260.

Alderson, J. C. 2007. The CEFR and the need for more research. *The Modern Language Journal, 91*(4), 659-663.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., and Tardieu, C. 2006. Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly, 3*(1), 3–30.

Alderson, J. C., and Huhta, A. 2005. The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22*(3), 301-320.

Buckendahl, C. W. 2005. Guest editor's introduction: Qualitative inquiries of participants' experiences with standard setting. *Applied Measurement in Education, 18*(3), 219-221.

Cizek, G. J., and Bunch, M. 2007. *Standard setting: A guide to establishing and evaluating performance standards on tests*. London: Sage Publications.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. 2009. Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. A Manual. Retrieved 15/02/2009, from http://www.coe.int/t/dg4/linguistic/manuel1_en.asp

Deville, C., and Chalhoub-Deville, M. 2011. Accountability-assessment under No Child Left Behind: Agenda, practice, and future. *Language Testing, 28*(3), 307-321.

Figueras, N., and Noijons, J. (Eds.). 2009. *Linking to the CEFR levels: Research perspectives*. Arnhem: CITO.

Fulcher, G. 2004. Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly, 1*(4), 253-266.

Garcia Gomez, P., Noah, A., Schedl, M., Wright, C., and Yolkut, A. 2007. Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing, 24*(3), 417-444.

Hambleton, R. K. 2001. Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89 -116). Mahwah, N.J.: Lawrence Erlbaum Associates.

Hambleton, R. K., and Pitoniak, M. J. 2006. Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433-470). Westport, CT: Praeger Publishers.

Hasselgreen, A. 2005. Assessing the language of young learners. *Language Testing, 22*(3), 337-354.

Kane, M. 1994. Validating the Performance Standards Associated with Passing Scores. *Review of Educational Research, 64*(3), 425-461.

Kane, M. 2012. Validating score interpretations and uses. *Language Testing, 29*(1), 3-17.

Kantarcioglu, E., and Papageorgiou, S. 2011. Benchmarking and standards in language tests. In B. O' Sullivan (Ed.), *Language Testing: Theories and Practices* (pp. 94-110). Basingstoke: Palgrave.

Kantarcioglu, E., Thomas, C., O'Dwyer, J., and O' Sullivan, B. 2010. Benchmarking a high-stakes proficiency exam: the COPE linking project. In W. Martyniuk (Ed.), *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's draft manual* (pp. 102-116).

Linacre, J. M. 1994. *Many-facet Rasch measurement* (2nd ed.). Chicago: MESA Press.

Little, D. 2006. The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching, 39*(3), 167-190.

Martyniuk, W. (Ed.). 2010. *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's draft manual*. Cambridge: Cambridge University Press.

North, B. 2000. *The development of a common framework scale of language proficiency*. New York: Peter Lang.

North, B., and Schneider, G. 1998. Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217-262.

Papageorgiou, S. 2010a. Investigating the decision-making process of standard setting participants. *Language Testing, 27*(2), 261-282.

Papageorgiou, S. 2010b. *Setting cut scores on the Common European Framework of Reference for the Michigan English Test.* (Technical Report). Ann Arbor:

University of Michigan. Retrieved 09/09/2011, from http://www.cambridgemichigan.org/sites/default/files/resources/MET_StandardSetting.pdf.

Plake, B. S. 2008. Standard setters: Stand up and take a stand! *Educational measurement: Issues and Practice, 27*(1), 3-9.

Shohamy, E., and McNamara, T. 2009. Language tests for citizenship, immigration, and asylum. *Language Assessment Quarterly, 6*(1), 1-5.

Tannenbaum, R. J., and Katz, I. R. 2013. Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol 3. Testing and assessment in school psychology and education* (pp. 455-477). Washington, DC: American Psychological Association.

Van Ek, J. A., and Trim, J. L. M. 1991. *Waystage 1990*. Cambridge: Cambridge University Press.

Van Ek, J. A., and Trim, J. L. M. 1998. *Threshold 1990*. Cambridge: Cambridge University Press.

Van Ek, J. A., and Trim, J. L. M. 2001. *Vantage*. Cambridge: Cambridge University Press.

Webb, N. L. 2007. Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*(1), 7-25.

Wilkins, D. A. 1976. *Notional syllabuses*. Oxford: Oxford University Press.

Zieky, M. J., and Perie, M. 2006. A primer on setting cutscores on tests of educational achievement. Retrieved 08/08/2011, from http://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf

Received: 26 November 2013

Accepted: 27 February 2014

Cite this article as:

Papageorgiou, S. 2014. "Issue in aligning assessments to the Common European Framework of Reference". *Language Value* 6 (1), 15-27. Jaume I University ePress: Castelló, Spain. http://www.e-revistes.uji.es/languagevalue. DOI**:** http://dx.doi.org/10.6035/LanguageV.2014.6.3

ISSN 1989-7103