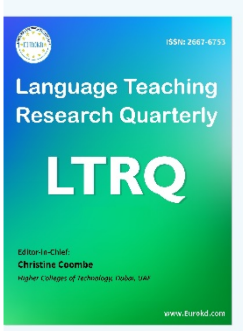

ISSN: 2667-6753

Language Teaching Research Quarterly

LTRQ

Editor-In-Chief:
Christine Coombe
Higher Colleges of Technology, Dubai, UAE

www.Eurokd.com

Language Teaching Research Quarterly

2022, Vol. 29, 57–64

EUROPEAN KNOWLEDGE DEVELOPMENT Ltd.
EUROKD

Still Deluded by Artifices? The Role of the Common European Framework of Reference in Facilitating Test Score Interpretation

Spiros Papageorgiou

Managing Senior Research Scientist, Center for Language Education and Assessment Research (CLEAR), Educational Testing Service, United States

Received *22 November 2021* ***Accepted*** *20 April 2022*

Abstract

The introduction of the CEFR was welcomed by researchers and practitioners in language education, given its potential for increasing transparency of test results and the meaningfulness of test scores. In this paper, I reflect on Glenn Fulcher's (2004, 2016) critical take on the use of the CEFR in the context of mapping (linking or aligning) test scores to the CEFR proficiency levels and the implications for score interpretation and use. I argue that although mapping test scores to the CEFR levels can enhance score interpretation, the field of language assessment needs to address misinterpretations of score mapping as sufficient evidence of quality of test design or comparability of scores of different tests.

Keywords: *Common European Framework of Reference, Test Score Interpretation, Score Mapping, Standard Setting*

Introduction

In the past two decades, the field of language assessment has seen a growing body of published research on mapping (aligning or linking) test scores to language proficiency levels of language frameworks, in particular those levels presented in the Common European Framework of Reference (CEFR). The introduction of the CEFR was welcomed by researchers and practitioners in language education, given its potential for increasing transparency of test results

spapageorgiou@ets.org

doi:10.32038/ltrq.2022.29.04

and the meaningfulness of test scores (Alderson, 2007; Kane, 2012). The CEFR was published by the Council of Europe to provide "a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe" (Council of Europe, 2001, p. 1). The impact of the CEFR has been particularly noticeable in the field of language assessment. In fact, Little (2007, in press) points out that such impact far outweighs the impact the CEFR had on curriculum design and pedagogy.

Despite the potential for positive impact, there has been considerable criticism on the uses of the CEFR as a policy document (McNamara, 2006). Glenn Fulcher's paper *Deluded by Artifices? The Common European Framework and Harmonization* (Fulcher, 2004) was perhaps the first elaborate, critical account of the CEFR as a tool for governments to implement harmonization across educational systems or for testing agencies to gain wider recognition for their tests. In this paper I discuss the process of mapping test scores to the CEFR proficiency levels and the implications for score interpretation and use. I argue that although mapping test scores to the CEFR levels can enhance score interpretation, the field of language assessment needs to address misinterpretations of score mapping as sufficient evidence of quality of test design or comparability of scores of different tests.

The Development of the CEFR and its Proficiency Scales

The Council of Europe published a number of documents in the 1970s that have been influential in second language teaching. Such documents include the notional-functional syllabus by Wilkins (1976) that describes what a learner communicates through language and three ascending levels describing language achievement: Waystage (Van Ek & Trim, 1991), Threshold (Van Ek & Trim, 1998) and Vantage (Van Ek & Trim, 2001). The CEFR emerged from this ongoing work of the Council of Europe, as well as North's research (North, 2000), as a publication in 2001 (Council of Europe, 2001). It contains dozens of language proficiency scales, describing language activities and competences at six main levels: A1 (the lowest) through A2, B1, B2, C1 and C2 (the highest). These six "criterion" levels are complimented in some scales by intermediate 'plus' levels, e.g. A2+, B1+, and B2+. The CEFR scales comprise statements called descriptors, which were designed following an action-oriented approach, where language users are seen as members of a society who have tasks to accomplish, including those that are not language-related.

The CEFR proficiency scales and performance descriptors were developed based on both quantitative and qualitative methodologies during a large-scale research project reported in North and Schneider (1998) and in more detail in North (2000). An initial pool of forty-one proficiency scales with their constituent descriptors was created based on existing ones, such as the ACTFL Proficiency Guidelines (for a detailed list, see Council of Europe, 2001, pp. 224-225). In the next, qualitative phase, the scales were refined though consultations with teachers representing all educational sectors in Switzerland. The refined scales and descriptors underwent quantitative analysis by asking teachers to use them to rate the performance of their students as well as selected student performances provided by the project team in video format. Using the many-facet Rasch model (Linacre, 1994), the descriptors were then calibrated and placed at different proficiency levels that subsequently formed the CEFR levels.

Although the CEFR is mostly known for these proficiency scales, it also contains rich information on language learning and assessment. For example, Chapter 5 discusses general competences and communicative language competences. Chapter 7 analyzes the role of tasks, both real-life and classroom ones, in language learning and teaching. Chapter 9 is concerned with topics related to assessment, in particular the various assessment purposes and types of assessments. The CEFR document also contains a number of appendices which provide supplementary material, including a discussion of technical issues specific to the development of language proficiency scales and performance descriptors. A companion volume was published in 2020 (Council of Europe, 2020). This volume did not change the number of proficiency levels, but it added new descriptors to further illustrate the levels and the various language activities and competencies.

The Role of the CEFR: Language Proficiency Standard, Framework, or Model?

The terms *framework* and *standard* are often used in the language testing literature to define the CEFR. However, neither term seems to fully represent the role of the CEFR as a reference source for the development of curricula, textbooks, and tests. A standard might refer to a set of guidelines on which tests are constructed and evaluated as Alderson et al. (1995, p. 236) note. One such set of guidelines is the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). An example of how the Standards have helped evaluate language tests is through their use as the basis for the *ETS Standards for Quality and Fairness* (ETS, 2014). The ETS Standards are used throughout the process of design, development, and delivery of language tests, such as the TOEFL iBT test, and their ongoing auditing to help ensure technical quality, fairness, and usefulness of test scores. The term standard might also describe learning outcomes used to assess and report learner progress and achievement, typically in the form of behavioral scales of language proficiency (Brindley, 1998), and they are found not only in educational contexts, but also professional ones where language proficiency is a requirement for performing in the workplace. In the context of aviation English, for example, the International Civil Aviation Organization (ICAO) has set English language requirements for air traffic controllers and pilots (Alderson, 2010). Milanovic and Weir (2010) point out that the CEFR levels do not constitute standards in the strictest sense, but instead provide a useful frame of reference and a source of meta-discourse. However, the CEFR is commonly used as an external standard by governments and international agencies in order to set policy and language proficiency goals (Fulcher, 2016).

Another common term for referring to the CEFR is the potentially misleading *framework*, which also appears in its title. According to Davidson and Fulcher (2007), documents such as the CEFR are in fact *models*, because they constitute a general description of language competence and demonstrate a theoretical understanding of language knowledge and use. A framework would then be used to select skills and abilities from a model that are relevant for an assessment context. For example, a framework would help test developers move from a model to a test blueprint that determines the content and format of a language test and provides the rationale for construct operationalization (see for example Papageorgiou et al., 2021).

Despite possible confusion about the role of the CEFR, its publication has been recognized as the "most significant recent event on the language education scene in Europe" (Alderson, 2005, p. 257), and its impact has been felt beyond the continent's borders, as language examination providers inside and outside Europe follow various methodologies to map the scores of their tests to the CEFR levels, as reported in several case studies in Figueras and Noijons (2009) and Martyniuk (2010).

Mapping Test Scores to the CEFR Levels

Because decisions based on test scores can have important consequences for students, teachers, and institutions, it is critical that test results be communicated in ways as transparent and meaningful as possible (Tannenbaum, 2019). However, test scores typically do not convey direct information about what test takers actually know and are able to do. Mapping test scores to the CEFR levels aims to facilitate the interpretation of these scores (Tannenbaum & Cho, 2014). When levels such as those in the CEFR are relevant to the constructs being measured by a particular test and widely known in the educational contexts where the test is administered, then mapping can often make the interpretation of test scores meaningful to the community familiar with these levels and descriptors (Powers et al., 2017). Ultimately, mapping is a claim about the interpretation of test scores in relation to external levels of language proficiency. To support such a claim, established procedures should be carefully implemented and multiple sources of evidence should be collected.

To help test providers follow robust procedures to map scores to the CEFR levels, the Council of Europe published the *Manual* in 2009 (Council or Europe, 2009). The mapping procedure in the Manual consists of three interconnected stages:

- Specification stage (construct congruence), which explores the extent to which the test adequately covers what is described in the CEFR.
- Standardization stage, which aims to set minimum scores (cut scores) to classify test takers in levels of performance following standard setting methodology (Cizek & Bunch, 2007).
- Empirical validation, which aims to provide evidence supporting the standard setting results from alternative sources (teacher ratings, student self-assessments, scores on other tests, etc.)

Two collections of case studies (Figueras & Noijons, 2009; Martyniuk 2010) and various published reports (e.g., Baron & Papageorgiou, 2014; Lim et al., 2013; Papageorgiou et al., 2015) share the experience of various test providers and researchers in mapping test scores to the CEFR levels, which varies in several ways. For example, some case studies in Martyniuk (2010) report on the alignment of a single assessment to the CEFR, whereas others deal with suites of examinations or multinational, large-scale research projects. Despite the differences found in terms of test content, score use, and methodology, common threads can be found in these publications. One thread is that score mapping to the CEFR might not be straightforward because, by design, its description of what learners are expected to do is under-specified to allow for a wider application. This intended under-specification might make the mapping process for specific groups of test takers challenging, for example young learners (see Papageorgiou & Baron, 2017). Despite several issues faced during the score-mapping process, a positive effect reported in some studies is raising awareness of important assessment design issues in contexts where local tests were developed and used. For example, following the score-mapping project of the COPE test (Kantarcioglu et al., 2010), revisions were made to the writing prompts, and features that differentiated passages and items across levels of language ability were included in the test specifications.

Challenges in Score Interpretation

As mentioned earlier, the mapping of test scores to the CEFR levels allows for increasing transparency of test results and the meaningfulness of test scores (Alderson, 2007; Kane, 2012). However, there are some noticeable misinterpretations of test scores that have been mapped to the CEFR levels. One such misinterpretation relates to content equivalence or interchangeability of test scores. Learners, teachers, and score users might view tests whose scores have been mapped to the same CEFR levels as equivalent in terms of difficulty or content coverage when this should not be the case (see Council of Europe, 2009, p. 11). For example, achieving CEFR Level B1 on a general proficiency test intended for young learners and a test intended for professional purposes does not mean that the scores on these two tests can be interpreted in the same way because test purpose, test content, and the target test taking population are notably different. For this reason, empirically derived, test specific performance levels and descriptors might need to be designed for a given test, for example by following a scale anchoring methodology (e.g., Powers et al., 2017). Such levels and descriptors can be provided in addition to information about mapping to the CEFR levels, so that score interpretation is relevant to the construct operationalized by the test.

A second misinterpretation of test scores that have been mapped to the CEFR levels is that mapping is sufficient evidence of the validity of these scores. The *Manual* strongly emphasizes the quality of a test as a prerequisite for score mapping, otherwise the mapping effort is "a wasted enterprise" (Council of Europe, 2009, p. 90). For example, it is pointless to set a cut score in relation to a specific CEFR level for a test with low internal consistency, as the measurement error associated with this cut score will be large. It is also unlikely that tests that rely primarily on decontextualized, discrete-point questions can demonstrate sufficient construct congruence with the CEFR and its proficiency levels, which follow a task-based, action-oriented approach. A possible reason why score mapping is misinterpreted as sufficient evidence of the validity of scores is because, as Fulcher (2016) points out, *validation* becomes synonymous with *recognition* when decision makers set requirements using the CEFR levels. If, for example, universities require international students to demonstrate proficiency at CEFR B2 level when applying for admission, then the test provider has a commercial interest to gain wider recognition by mapping test scores to the CEFR levels whether or not the actual abilities expected at B2 level have been adequately measured (Fulcher, 2004).

A third misinterpretation relates to how test providers interpret the CEFR levels. In his review of a test of academic language proficiency Green (2018) notes discrepancies across test providers to the point that "one [testing] agency's B2 may be another's A2/B1: the outcomes of the different linking approaches do not support each other closely and do not provide convincing mutual validation" (p. 71). For example, a score concordance study conducted by ETS found that TOEFL iBT total test scores of 60 to 78 correspond to an IELTS total band level of 6 (ETS, 2010). Although IELTS Band 6 is mapped to Level B2 (Lim et al., 2013), the TOEFL iBT test score range for the same CEFR level is 72–94 (Papageorgiou et al., 2015). Therefore, based on the corresponding TOEFL iBT scores, the lower end of IELTS Band 6 could be interpreted as CEFR Level B1 rather than Level B2. However, as Papageorgiou et al. (2015) point out, there is no authorized interpretation of the CEFR levels. Although, as North (2014) notes, the CEFR was intended as a tool for reference and consultation, its use for setting score requirements inevitably assumes a common interpretation of the CEFR levels across different tests.

The misinterpretations discussed in this section have implications for score-based decision-making in both local and global language education contexts. First, stakeholders who use test scores to make important decisions (e.g., placement into language classrooms and admission to university programs) might ignore the extent to which a language test is appropriate for a given purpose, and instead use any test just because its scores are mapped to the CEFR levels. Second, there might be little attention to validity evidence supporting the interpretation and use of test scores, leading to potentially harmful decisions, for example placing students in language classes that are too difficult or too easy.

Conclusion

Figueras (2012) attributes the success of the CEFR to its function as a common currency in terms of terminology and levels of attainment. The action-oriented approach of the CEFR offers a way to link learning, teaching, and assessment to real-life use of the language, presenting language proficiency through defined levels for language activities, tasks, and competences. Although the theoretical underpinnings of the CEFR are weak (Alderson, 2007), Fulcher (2004) acknowledges the usefulness of can-do statements such as those in the CEFR for conveying "a generalizable meaning of test scores to users, in terms of what a test taker with a particular score on a given test may typically be able to do" (p. 264). Fulcher's (2004, 1016) critical perspective into the use of the CEFR for a variety of purposes in the context of language education has helped those involved in score-mapping studies to critically reflect on how such mapping should be conducted so that it can support score interpretation in a meaningful way. In fact, recent applications of the process described in the Manual can even be seen in studies mapping scores of large-scale, international tests to local proficiency levels rather than the CEFR (Dunlea et al., 2019; Papageorgiou et al., 2019). Such localized efforts underscore the importance of considering contextual factors and how test scores might be interpreted or misinterpreted in a specific educational context as a result of the score mapping.

References

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal, 91*(4), 659-663. https://doi.org/10.1111/j.1540-4781.2007.00627_4.x

Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing, 27*(1), 51-72. https://doi.org/10.1177/0265532209347196

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/open-access-files.html

Baron, P. A., & Papageorgiou, S. (2014). *Mapping the TOEFL® Primary™ Test onto the Common European Framework of Reference* (ETS Research Memorandum No. RM-14-05). Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RM-14-05.pdf

Brindley, G. (1998). Describing language development? Rating scales and second language acquisition. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 112-140). Cambridge University Press.

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

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

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A Manual*. Retrieved from http://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2d

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume*. Council of Europe.

Davidson, F., & Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching, 40*(3), 231-241. https://doi.org/10.1017/S0261444807004351

Dunlea, J. Spiby, R., Wu S., Zhang, J. & Cheng, M.M (2019). *China's Standards of English Language Ability (CSE): Linking UK exams to the CSE* (Research Report No. VS/2019/0003). Retrieved from https://www.britishcouncil.org/sites/default/files/linking_cse_to_uk_exams_5_0.pdf

ETS. (2010). *Comparing TOEFL® and IELTS™ total scores*. Retrieved from https://www.ets.org/toefl/institutions/scores/compare

ETS. (2014). *ETS Standards for quality and fairness*. Retrieved from https://www.ets.org/s/about/pdf/standards.pdf

Figueras, N. (2012). The impact of the CEFR. *ELT Journal, 66*(4), 477-485. https://doi.org/10.1093/elt/ccs037

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

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly, 1*(4), 253-266. https://doi.org/10.1207/s15434311laq0104_4

Fulcher, G. (2016). Standards and frameworks. In D. Tsagari & J. Banerjee (Eds.), *Handbook of Second Language Assessment* (pp. 29-44). De Gruyter Mouton.

Green, A. (2018). Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly*, *15*(1), 59-74. https://doi.org/10.1080/15434303.2017.1350685

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3-17. https://doi.org/10.1177/0265532211417210

Kantarcioglu, E., Thomas, C., O'Dwyer, J., & 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). Cambridge University Press.

Lim, G. S., Geranpayeh, A., Khalifa, H., & Buckendahl, C. W. (2013). Standard setting to an international reference framework: Implications for theory and practice. *International Journal of Testing*, *13*(1), 32-49. https://doi.org/10.1080/15305058.2012.678526

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

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal, 91*(4), 645-655. https://doi.org/10.1111/j.1540-4781.2007.00627_2.x

Little, D. (in press). Introduction to Part 1: The *Common European Framework of Reference for Languages*: past, present and future. In D. Little & N. Figueras (Eds.), *Reflecting on the Common European Framework of Reference for Languages and its Companion Volume*. Multilingual Matters.

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 University Press.

McNamara, T. (2006). Validity in language testing: the challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3*(1), 31-51. https://doi.org/10.1207/s15434311laq0301_3

Milanovic, M., & Weir, C. J. (2010). Series Editors' note. 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. viii-xx). Cambridge University Press.

North, B. (2000). *The development of a common framework scale of language proficiency*. Peter Lang.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217-262. https://doi.org/10.1177/026553229801500204

North, B. (2014). Putting the Common European Framework of Reference to good use. *Language Teaching, 47*(2), 228-249. doi:10.1017/S0261444811000206

Papageorgiou, S., & Baron, P. A. (2017). Using the Common European Framework of Reference for young learners' English language proficiency assessments. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 136-152). New York: Routledge. https://doi.org/10.4324/9781315674391

Papageorgiou, S., Davis, L., Norris, J. M., Garcia Gomez, P., Manna, V. F., & Monfils, L. (2021). *Design framework for the* TOEFL® Essentials™ *test 2021* (Research Memorandum No. RM-21-03). Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RM-21-03.pdf

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFLiBT® test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum No. RM-15-06). Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RM-15-06.pdf

Papageorgiou, S., Wu, S., Hsieh, C.-N., Tannenbaum, R. J., & Cheng, M. M. (2019). *Mapping the* TOEFL iBT® *test scores to China's Standards of English Language Ability: Implications for score interpretation and use* (Research Report No. TOEFL-RR-89). Educational Testing Service. https://doi.org/10.1002/ets2.12281

Powers, D., Schedl, M., & Papageorgiou, S. (2017). Facilitating the interpretation of English language proficiency scores: Combining scale anchoring and test score mapping methodologies. *Language Testing, 34*(2), 175-195. https://doi.org/10.1177/0265532215623582

Tannenbaum, R. J. (2019). Validity aspects of score reporting. In. D. Zapata-Rivera (Ed.), *Score reporting research and applications* (pp. 9-18). Routledge.

Tannenbaum, R. J., & Cho, Y. (2014). Criteria for evaluating standard-setting approaches to map English languagetest scores to frameworks of English language proficiency. *Language Assessment Quarterly*, *11*(3), 233-249. https://doi.org/10.1080/15434303.2013.869815

Van Ek, J. A., & Trim, J. L. M. (1991). *Waystage 1990*. Cambridge University Press.

Van Ek, J. A., & Trim, J. L. M. (1998). *Threshold 1990*. Cambridge University Press.

Van Ek, J. A., & Trim, J. L. M. (2001). *Vantage*. Cambridge University Press.

Acknowledgments

Not applicable.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

Rights and Permissions

Open Access