The reification of the Common European Framework of Reference (CEFR) and effect-driven testing

Glenn Fulcher

University of Leicester

Abstract

Standards-based educational systems, in which assessment is the fundamental tool for accountability, are increasingly being adopted by governments as a cheap and simple solution for dealing with perceived threats to global economic competiveness (Brindley 2008). Centralizing authorities see control and standardization as essential for economic prosperity by tackling the perceived failings of education and training in comparison with competitors. Despite the original intentions of the authors, the Common European Framework of Reference (CEFR) is now being adopted as a tool in standards-based education in Europe. As other countries and transnational organizations seek collectivist solutions to international problems, its adoption beyond European borders testifies to its usefulness in centralized language education policy. This paper argues that such use requires the unjustified reification of the CEFR levels, leading to a naïve understanding of communication and language acquisition.

Keywords: language testing, CEFR, reification, rating scales, language acquisition

1. The reification of the CEFR

The CEFR is increasingly being used as a tool for harmonization of language teaching, learning and assessment (Fulcher 2004). It is therefore not surprising that the 'validation' of the CEFR is interpreted as institutional recognition (Trim 1996: 416-417). As recognition grows, we see emerging calls for "the policing of the CEFR levels" (Alderson 2007: 662, Bonnet 2007: 671) as part of the process of removing the principle of subsidiarity from the field of education, and moving toward "a common educational policy in language learning, teaching and asessment, *both at the EU level and beyond*" (Bonnet 2007: 672, emphasis added).

The rapid spread of the use of the CEFR across Europe and other parts of the world can be accounted for by the ease with which it can be used in standards-based assessment, and form the basis for policy areas such as immigration (Krumm 2007). Malone (2008: 225-226) compares the CEFR with the No Child Left Behind (NCLB) legislation in the United States in terms of its influence, despite the fact that the CEFR is not mandatory. As a policy tool for harmonization we have seen that "large-scale

operations like the CEFR may be manipulated unthinkingly by juggernaut-like centralizing institutions" (Davies 2008: 438), which are also using the CEFR to define required levels of achievement for school pupils (Alderson 2007: 662) as well as adult language learners (North 2007: 657). Once such large-scale operations are perceived as 'the system', Davies (2008: 438) adds that they have historically resulted in "reducing diversity and experimentation" in research and language pedagogy.

The indiscriminate exportation of the CEFR for use in standards-based education and assessment in non-European contexts, such as Hong Kong and Taiwan, shows that Trim was correct when he observed that "there will always be people who are trying to use it as an instrument of power...." (ibid., 282). The problem, as Fulcher (2008: 170) puts it, is:

> "It is a short step for policy makers, from 'the standared required for level X' to 'level X is the standard required for....' This illigitimate leap of reasoning is politically attractive, but hardly ever made explicit or supported by research."

For this step to take place, a framework has to undergo a process of reification, a process defined as "the propensity to convert an abstract concept into a hard entity" (Gould 1996: 27). The only significant survey undertaken on the use of the CEFR in Europe (Council of Europe 2005) adds substantially to the evidence that the scales are being interpreted as a statement of how language acquisition really takes place.

2. The CEFR scales

The assumption that the CEFR scales have been constructed on a principled analysis of language use within a range of domains, or a theory of second language acquisition, is mistaken. The scale descriptors were drawn from existing scales in many different testing systems from around the world, and were placed within the CEFR scales because teacher judgments of their difficulty could be scaled using multi-faceted Rasch (North 1996). The steps in the development process are as follows (Fulcher 2003: 107-113):

Phase 1

Step 1: Collection of 2000 descriptors from over 30 scales in use around the world.

Step 2: Classification of each descriptor according to categories of communicative language ability and writing additional descriptors to fill perceived gaps.

Phase 2

Step 3: Pairs of teachers are given sets of descriptors typed onto confetti like strips of paper and asked to sort them into categories.

Step 4: The same pairs are asked to comment on the "usefulness" and "relevance" of each descriptor for their students.

Step 5: Teachers are given the same sets of descriptors and asked to separate them into three levels: 'low', 'middle' and 'high', and then divide each of these into two categories to create the familiar six level scale.

Step 6: The descriptors most consistently placed in the same level of the scale are used to create overlapping 'questionnaires' of descriptors, with the overlap items operating as anchors.

Phase 3

Step 7: A rating scale is attached to each descriptor on the questionnaire.

Step 8: A group of teachers is asked to rate a small number of their learners from their classes on the rating scale for each of the descriptors on the questionnaire.

Step 9: This data is used to construct scales of unidimensional items using Rasch analysis, rejecting any items that misfit the Rasch model.

Step 10: Items that behave statistically differently across languages or sectors are identified and removed.

Step 11: Cut scores are established using difficulty estimates in order to achieve equidistant bands.

Phase 4

Step 12: Conduct the study again using a different group of teachers.

The selection of descriptors for the CEFR scales, and scale assembly, was psychometrically driven; or as North (1995) says, based entirely on a theory of measurement. The data in the scaling studies were intuitive teacher judgments rather than samples of performance. What we see in the CEFR scales is therefore "essentially a-theoretical" (Fulcher 2003: 112), a critique which North and Schneider (1998: 242-243) admit to be the case. Since this analysis, it has been frequently repeated that the scales have no basis in theory or SLA research (Hulstijn 2007: 666).

These Frankenstein scales therefore need to be treated with great care. As collections of scaled proficiency descriptors it is not reasonable to expect them to relate to any specific communicative context, or even to provide a comprehensive (let alone exhaustive) description of any particular communicative language ability. Most

importantly, we cannot make the assumption that abilities do develop in the way implied by the hierarchical structure of the scales. The scaling methodology assumes that all descriptors define a statistically unidimensional scale, but it has long been known that the assumed linearity of such scales does not equate to how learners actually acquire language or communicative abilities (Fulcher 1996b, Hulstijn 2007, Meisel 1980). Statistical and psychological unidimensionality are not equivalent, as we have long been aware (Henning 1992). The pedagogic notion of "climbing the CEFR ladder" is therefore naïve in the extreme (Westhoff 2007: 678). Finally, post-hoc attempts to produce benchmark samples showing typical performance at levels inevitably fall prey to the same critique as similar ACTFL studies in the 1980s, that the system states purely analytic truths: "things are true by definition only" (Lantolf and Frawley 1985: 339), and these definitions are both circular and reductive (Fulcher 2008: 170-171).

The reification of the CEFR is therefore not theoretically justified. However, reification is a necessary step to imposing harmonization through the requirement that assessments and teaching is aligned to the CEFR as an external standard.

3. Alignment to standards vs. effect driven testing

The use of the CEFR for harmonization has commitment only to system, and not to effect (Davidson and Fulcher 2006, 2007: 232); harmonization needs reified models that exist independently of the effects of a test on its users and their needs, whereas a commitment to effect requires variable frameworks and tests for different user populations and needs.

This reading is in keeping with Trim (1996: 417), when he expressed the view that a *model* on the scale of the CEFR deliberately lacked the detail necessary for local decision making and action, which is rightly the domain of the practitioner. When the CEFR is seen merely as a *heuristic model* which may be used at the will of the practitioner (or not used, if it does not suit the practitioner's purpose), it may become a useful tool in the construction of tests or learning activities.

In this discussion, we have now moved from calling the CEFR a 'framework', to calling it a 'model'. In effect-driven testing the purpose of a model is to act as a source of ideas for the selection of constructs that are useful and relevant to the design of tests for specific purposes. It is impossible to test everything that a model contains, for it is intended to be an encyclopaedic taxonomy of what we know about language that is not tied to any context of use. Language testing, on the other hand, rightly prioritizes the

purpose for which we test, and makes context the driver of test design decisions. The context of language use is therefore critical, as it places limitations upon the legitimate inferences that we might draw from test scores, and restricts the range of decisions to which the score might be relevant. The constructs are articulated in a test *framework* that provides a theoretical rationale for the relevance of the constructs to the specific context, and the operationalization of these constructs is embodied in the test *specifications* (Chalhoub-Deville 1997, Fulcher 2004, Fulcher and Davidson 2007). The *model*, the *framework*, and *specifications*, are referred to as the three levels of architectural documentation in test design (Fulcher 2006: 5, Fulcher and Davidson 2009).

The CEFR is not a 'framework' in this sense. It is a high-level generic model. Yet, the term 'framework' in its title suggests that it is capable of generating test specifications, or being the medium by which existing tests and specifications can be compared. The clearest example of this fallacy lies in the assumption that by mapping the content of an existing test onto the content of the CEFR one can demonstrate 'linkage' between the two. In assuming that test specifications can be compared directly with a model (as in Alderson *et al*. 2006) the conclusion that a test specification will under-represent the model is as inevitable as the conclusion that the model does not contain the detail that is needed for an operational test specification. The two documents are at different levels of test architecture, and cannot be directly compared. This is why it is illogical to attempt to compare tests designed for different purposes through the medium of an encyclopaedic model. The illogicality of this position is made particularly clear in Kaftandjieva (2007), which is a call to establish a quantitative component to content linkages between actual tests and the CEFR. Kaftandjieva (2007: 35) argues that validity is a question of alignment, and that "the main goal of this linking exercise is to demonstrate compliance with a mandate." Acknowledging the problems that many have had with alignment studies (e.g. Alderson *et al.* 2006), Kaftandjieva (2007: 36-37) recommends matching test content with a standards document on the two dimensions of topic and cognitive demand. Applying this to a particular reading comprehension test, Kaftandjieva discovers that the match between the test and the CEFR reading descriptors is approximately 27%, which she accounts for by claiming that the CEFR descriptors are very general, and that the test is not long enough. She then argues that linkage can be improved "simply by adding a few more items based on short texts whose discourse type is expository, argumentative or instructive" (ibid., 40) and by

combining categories in the CEFR that judges cannot agree upon, such as "make straightforward inferences" and "interpret and integrate ideas and information".

Improving alignment by adding as many discourse types (along with a few items) that are mentioned in the CEFR merely leads test developers to believe that they are testing everything, for all purposes. The advice provided to achieve alignment is not burdened with any concern that the scores might be used to make decisions about the application of reading for actual purposes in the real world. Once again, the concern is entirely with harmonization – and harmonization interpreted as validity – rather than effect. Further, as harmonization is a matter of judgment, or 'social moderation' (North 2000), if linkage cannot be achieved immediately, judges are merely trained more thoroughly to produce the required judgments (Fulcher, 2010). This is a throwback to similar validity claims for the ACTFL Guidelines, the circularity of which is well documented (Chalhoub-Deville and Fulcher 2003, Fulcher 1996a, 2008).

Effect driven testing, on the other hand, is concerned with the probability that a score interpretation coincides with its corresponding meaning in the intended object of measurement, and that decisions made on the basis of interpretations are not applied beyond contexts or domains of reasonable extrapolation. Within effect driven testing, the design process is concerned with creating an interpretative argument (Kane 2006) that relates design decisions to intended effects, and when applied to proficiency scales the most general claim is that "the quality of performance required for level B is A", where A and B are not arbitrary because they are related to context through a validity argument. Yet, A and B are not constants, and so cannot be applied as universal solutions to all testing problems. Validity is located in the quality of an argument, not a measure of alignment.

Effect-driven testing therefore requires that test design principles explicitly link claims of score meaning directly to test performance and the intended universes of generalization and extrapolation through a validity argument. In performance testing the design of the rating (or scoring) procedure is often central to such an argument because this is where the construct definition is operationalized (Fulcher 2003: 89). Rating scales that are sensitive to test purpose operationalize what we know about the context of language use in ways that are not static in the same way that scaled proficiency descriptors are. Rather, they can be crafted to reflect the interactional competence of the speakers as argued by Kramsch (1986), which can only be achieved when descriptions of performance are context sensitive, socially specific, and inherently local in meaning

(Chalhoub-Deville 2003, He and Young 1998). It no longer makes sense to articulate rating scales in terms of scaled hierarchical proficiency descriptors that do not anchor performance in context (Fulcher and Davidson 2007: 98-100). As Krumm (2007: 667) would put it:

> "...in a world of social, cultural, and individual heterogeneity, one instrument and approach can neither address all situations and contexts nor meet all needs. Although the CEFR is not intended to be applied uniformly to everybody, in some cases it is applied in just such a fashion...."

In the next section we will therefore briefly analyse one set of CEFR proficiency descriptors in order to demonstrate the lack of context necessary for the operationalization of dynamic interactional competence in assessment.

4. The CEFR service encounter descriptions and scales

Service encounters are claimed to occur mostly at level B1. These include the abilities to:

> "...make simple transactions in shops, post offices or banks; get simple information about travel; use public transport: buses, trains, and taxis, ask for basic information, ask and give directions, and buy tickets; ask for and provide everyday goods and services."

The public domain contexts in which these transactions may take place are provided in tables laid out in the CEFR (Council of Europe 2001: 48-49) taxonomy. Goods and services are thrown together, and there is no distinction between qualitatively different transactions (McCarthy and Carter 1994: 63, Ylänne-McEwen 2004: 518-519). Still less is there any attempt to distinguish between purchasing goods, and obtaining services that are "less tangible" (Coupland 1983: 464-465). Any section or item from this unstructured, incomprehensive list, may (or may not) be relevant to language use in a particular domain. It is therefore not surprising that the CEFR does not suggest any tasks that might be associated with transactional language use. Rather, users of the CEFR are asked to consider what task types might be relevant to a given context (Council of Europe 2001: 54).

As there are no tasks listed within the CEFR, it is obvious that there can be no performance conditions, defined as: "specific conditions that give us the purpose of

communication, setting/place, audience, topic, time constraints, length of task, assistance allowed, etc." (Pawlikowska-Smith 2000: ix). Rather, the CEFR invites the reader to consider how the physical conditions, number of interlocutors and time pressures, will impact on what the learner has to do (Council of Europe 2001: 50).

The list provides nothing that can act as a framework in test development, but its vagueness is also an advantage in that when used as a heuristic at the beginning of test development it provides no constraints whatsoever upon the test designers (Davidson and Fulcher 2007).

Moving from contexts to level descriptors for transactions, Table 1 below reproduces the CEFR illustrative scale.

Table 1. Illustrative scale for Transactions

	TRANSACTIONS TO OBTAIN GOODS AND SERVICES
C2	*As B2*
C1	*As B2*
B2	*Can cope linguistically to negotiate a solution to a dispute like an undeserved traffic ticket, financial responsibility for damage in a flat, for blame regarding an accident.* *Can outline a case for compensation, using persuasive language to demand satisfaction and state clearly the limits to any concession he/she is prepared to make.*
	Can explain a problem which has arisen and make it clear that the provider of the service/customer must make a concession.
B1	*Can deal with most transactions likely to arise whilst travelling, arranging travel or accommodation, or dealing with authorities during a foreign visit.* *Can cope with less routine situations in shops, post offices, banks, e.g. returning an unsatisfactory purchase. Can make a complaint.* *Can deal with most situations likely to arise when making travel arrangements through an agent or when actually travelling, e.g. asking passenger where to get off for an unfamiliar destination.*
A2	*Can deal with common aspects of everyday living such as travel, lodgings, eating and shopping.* *Can get all the information needed from a tourist office, as long as it is of a straightforward, non-specialised nature.*
	Can ask for and provide everyday goods and services. *Can get simple information about travel, use public transport: buses, trains, and taxis, ask and give directions, and buy tickets.* *Can ask about things and make simple transactions in shops, post offices or banks.* *Can give and receive information about quantities, numbers, prices, etc.* *Can make simple purchases by stating what is wanted and asking the price.* *Can order a meal.*
A1	*Can ask people for things and give people things.* *Can handle numbers, quantities, cost and time.*

We immediately face a number of problems with this scale and its descriptors that stem directly from the way in which it was constructed. Some descriptors refer to specific situations, while others do not. Level B2, for example, refers to getting a traffic ticket, damaging property, and dealing with being blamed for an accident. When a

context of language use is mentioned, it is not necessarily referred to in other descriptors. Dealing with travel agents is specifically mentioned in Level B1, but not at other levels, despite references to travel. We are therefore left with the question of whether 'dealing with travel agents' is something that is suddenly possible at level B1. Participant roles are mixed within the same level. At A2 for example, the leaner can "ask for and provide" goods and services. This seems to imply that an A2 learner would be able to function as a shop keeper as well as purchase items from a shop. At level B2 would this mean that a learner could explain to a client how to seek compensation, as well as ask for compensation as a customer? The distinction between levels is unclear, with descriptors referring to the vague concept of 'complexity' at each level. At level B1 learners can deal with "most" transactions, as well as "less routine" situations. But there is no definition of "less", "more" and "most". A2 is characterized by "common", "everyday", "simple" and "straightforward" transactions, but we are not told what these might be.

Just as the context for interaction is an unstructured list, the descriptors on the scale do not represent a linguistic or discourse analysis of the target domain. The descriptors are assembled in the way they are only by virtue of the fact that the perception of their difficulty is amenable to statistical manipulation.

5. Conclusion

The CEFR is incapable of addressing specific contexts of language use, but this is also its greatest strength. Its flexibility lies precisely in that fact that it is a static set of proficiency level descriptors that bear little relationship to transactional communication in the real world. The CEFR can therefore act as a heuristic that may aid test designers in initial planning for a new. However, it is not possible to generate tasks from the CEFR, to define the test construct without additional analysis of the context of language use, or to score test takers using the CEFR scales. These tasks remain in the domain of the professional practitioner. It also follows that once reification is rejected, we must also object to its use as a tool in standards-based assessment for language education in Europe, and beyond.

References

Alderson J.C., N. Figueras, H. Kuijper, G. Nold, S. Takala, and C. Tardieu (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. (2007). The CEFR and the need for more research. *The Modern Language Journal* 91/4: 659-663.

Bonnet G. (2007). The CEFR and education policies in Europe. *The Modern Language Journal* 91/4: 669-672.

Brindley G. (2008). Educational reform and Language testing. In E. Shohamy (ed.), *Language testing and assessment*. Encyclopedia of Language and Education, Vol 7. New York: Springer, 365-378.

Chalhoub-Deville M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing* 14/1: 3-22.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing* 20/4: 369 – 383.

Chalhoub-Deville, M. and G. Fulcher (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals* 36/4: 498–506.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching and assessment*. Cambridge: Cambridge University Press. Available online at: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf.

Council of Europe (2005). Survey on the use of the Common European Framework of Reference for Languages (CEFR): Synthesis of results. Retrieved 8 November, 2007: http://www.coe.int/t/dg4/linguistic/Source/Surveyresults.pdf.

Coupland N. (1983). Patterns of encounter management: Further arguments for discourse variables. *Language in Society* 12: 459-476.

Davidson F. and G. Fulcher (2006). Flexibility is proof of a good 'framework'. *Guardian Weekly*, 17th November. Retrieved 5 November, 2007: http://education.guardian.co.uk/tefl/viewfromabroad/story/0,,1950501,00.html.

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

Davies A. (2008). Ethics and professionalism. In E. Shohamy (ed.), *Language testing and assessment*. Vol. 7. Encyclopedia of Language and Education. New York: Springer, 429-443.

Fulcher G. (1996a). Invalidating validity claims for the ACTFL oral rating scale. *System* 24/2: 163-172.

Fulcher G. (1996b). 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/Pearson.

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

Fulcher G. (2008). Criteria for evaluating language quality. In E. Shohamy (ed.), *Language testing and assessment*. Encyclopedia of language and education, Vol 7. Amsterdam: Springer, 157-176

Fulcher G. (2006). Test architecture. *Foreign Language Education Research* 12: 1-22.

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.

Fulcher G. and F. Davidson (2007). *Language testing and assessment*. London and New York: Routledge.

Fulcher G. and F. Davidson (2009). Test architecture, test retrofit. *Language Testing*. 26/1: 123 – 144.

Gould, S.J. (1996). *The mismeasure of man*. London: Penguin.

He A.W. and R. Young (1998) Language proficiency interviews: a discourse approach. In R. Young and A.W. He (eds), *Talking and testing: Discourse approaches to the assessment of oral proficiency*. Amsterdam: John Benjamins.

Henning G. (1992). Dimensionality and construct validity of language tests. *Language Testing* 9/1: 1-11.

Hulstijn J. A. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal* 91/4: 663-667.

Kaftandjieva F. (2007). Quantifying the quality of linkage between language examinations and the CEF. In C. Carlsen, and E. Moe (eds), *A human touch to language testing*. Oslo: Novus Press.

Kane M.T. (2006). Validation. In R.L. Brennan (ed.) *Educational measurement*. Fourth Edition. New York: National Council on Measurement in Education and Praeger Publishers.

Kramsch C. (1986). From language proficiency to interactional competence. *The Modern Language Journal* 70/4: 366-372.

Krumm H-J. (2007). Profiles instead of levels: The CEFR and its (ab)uses in the context of migration. *The Modern Language Journal* 91/4: 667-669.

Lantolf J.P. and W. Frawley (1985). Oral proficiency testing: A critical analysis. *The Modern Language Journal* 69/4: 337-345.

Malone M. (2008). Training in language assessment. In E. Shohamy (ed.), *Language testing and assessment*. Vol. 7. Encyclopedia of Language and Education. New York: Springer, 225-239.

McCarthy M. and R. Carter (1994). *Language as discourse: Perspectives for language teaching*. London: Longman.

Meisel J.M. (1980). Linguistic simplification. In S. Felix (ed.), *Second language development: Trends and issues*. Tubingen: Gunter Narr, 13-40.

North B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System* 23/4: 445-465.

North B. (1996). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (eds), *Current developments and alternatives in language assessment*. Proceedings of the LTRC 1996. Jyvaskyla: University of Jyvaskyla Press, 423-447.

North B. (2000). Linking language assessments: an example in a low stakes context. *System* 28/4: 555-577.

North B. (2007). The CEFR illustrative descriptor scales. *The Modern Language Journal* 91/4: 656 – 659.

North B. and G. Schneider (1998). Scaling descriptors for language proficiency scales. *Language Testing* 1/2: 217-263.

Pawlikowska-Smith G. (2000). *Canadian language benchmarks 2000*. Ontario: Centre for Canadian Language Benchmarks.

Trim J.L.M. (1996). The proposed Common European Framework for the description of language learning, teaching and assessment. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma, (eds), *Current developments and alternatives in language assessment*. Proceedings of the LTRC, 1996. Jyvaskyla: University of Jyvaskyla Press, 415-421.

Westhoff G. (2007). Challenges and opportunities of the CEFR for reimagining foreign language pedagogy. *The Modern Language Journal* 91/4: 676 – 679.

Ylänne-McEwen V. (2004). "Shifting alignment and negotiating sociality in travel agency discourse. *Discourse Studies* 6/4: 517-536.