General Language Proficiency Revisited: Current and Future Issues

Claudia Harsch

To cite this article: Claudia Harsch (2014) General Language Proficiency Revisited: Current and Future Issues, Language Assessment Quarterly, 11:2, 152-169, DOI: 10.1080/15434303.2014.902059

To link to this article: http://dx.doi.org/10.1080/15434303.2014.902059

Published online: 21 May 2014.
General Language Proficiency Revisited: Current and Future Issues

Claudia Harsch

University of Warwick, Coventry, United Kingdom

This article explores a number of key issues that emerged during the panel discussion that followed the General Language Proficiency Symposium at the Language Testing Forum (LTF) 2010, celebrating the 30th anniversary of the LTF. The key issues that emerged during the discussion should be of interest to a wider audience, as they express current issues and concerns in the testing community in light of 30 years of research and development. These issues were whether language proficiency can be seen as unitary or divisible; the role and use of the Common European Framework (CEF) proficiency scales and levels when it comes to reporting test scores; the issue of test equivalence between high-stakes tests, also in relation to the CEF; and a general demand for developing assessment literacy among test users and stakeholders. Based on a summary of the LTF discussion, the article provides a state-of-the-art review of relevant research that addresses the aforementioned listed key issues, followed by a future agenda for researching the reviewed areas.

INTRODUCTION

This article explores a number of key issues that emerged during the panel discussion that followed the General Language Proficiency Symposium at the Language Testing Forum (LTF) 2010. The LTF 2010 had the theme “Issues in Language Testing Revisited,” celebrating the 30th anniversary of the LTF and revisiting the three main issues that had emerged from the first LTF in 1980: General Language Proficiency, Testing for Specific Purposes, and Communicative Language Testing. Each theme was addressed in a symposium, which was followed by a panel discussion. In what follows I take the key issues emerging from the General Language Proficiency panel discussion at the LTF 2010 as starting point. Although these key issues may give the article a certain, restricted focus, they nevertheless should be of interest to a wider audience as they express current issues and concerns in the testing community in light of 30 years of research and development. I review the current state of research for each of these selected key issues in turn in order to offer a better understanding about how these issues and challenges are currently being addressed. Based on the review of and insights from recent as well as ongoing research projects, I propose an up-to-date future agenda for researching the reviewed areas, offering new perspectives into possible next steps to move the field further.
The starting point of the LTF panel discussion was, of course, the notion of general language proficiency and whether it can be seen as unitary or divisible. The field has moved on since Oller’s (1979) claim of a global language factor, and nowadays a range of issues around the multidimensionality of language proficiency are being discussed and researched, targeting construct-related, developmental-oriented, and psychometric questions. Related to the conceptualisation of language proficiency, the panel discussion then moved on to its main concern, that is, what role the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) can play in testing language proficiency.\(^2\) Many of the concerns around the CEFR expressed at the LTF 2010 were related to the nature and divisibility of language proficiency, such as the conceptualisation and validation of different aspects of proficiency in the CEFR scale system, or the level of abstraction needed when reporting test scores for different test purposes. This article provides a state-of-the-art review of current projects researching the CEFR scales, their foundations, and their applicability to testing contexts. One aspect of employing the CEFR in testing contexts received particular attention during the LTF discussion: linking tests and exams to the CEFR, thus revisiting the question, “Is my B1 your B1?” Issues around CEFR alignment were heatedly debated at the LTF because they have major social, political, and economic implications in Europe and beyond. The majority of the international English proficiency test providers have linked their exams to the CEFR, despite problems and challenges that will be reviewed in detail in the article. Linkage to the CEFR, however, does not mean the tests themselves measure comparable aspects or can be treated as being equivalent. Yet this is often (wrongly) assumed, for example, for immigration or university admission purposes where administrators require simplified comparison tables, often reducing score profiles to one global score. Here, assessment literacy and education of stakeholders and test users are urgently needed to promulgate core knowledge in the field of language testing. Assessment literacy was the “red thread” throughout the LTF panel discussion, due to its potential of addressing many of the issues around language test use, which emerged during the discussion. Hence, the final aspect being reviewed in this article is dedicated to the current state of research on assessment literacy.

**GENERAL LANGUAGE PROFICIENCY—UNITARY OR DIVISIBLE?**

The question whether language proficiency can be conceptualised as unitary or divisible relates back to the 1970s, to Oller’s (1979) claim of a global language factor. The global language factor claim was rebutted by several researchers (e.g., Cummins, 1979) and by Oller himself (Oller, 1983),\(^3\) as Bachman pointed out in a recent interview (Chen, 2011). It emerged during the LTF discussion that this question was nowadays regarded as a “nonquestion,” as language proficiency can be conceptualised as unitary and divisible, depending on the level of abstraction and the purpose of the assessment and score reporting. This view is supported by recent research, for

---

\(^2\)The particular attention during the panel discussion on issues related to the CEFR may add to the focus of this article on research in contexts where the CEFR has played a dominant role (i.e., Europe and North America). Although the widening spread of language testing practices across the world was not the focus of discussion at the LTF and hence is not discussed here, this does not imply that research and test development in other parts of the world were not recognized by the author.

\(^3\)Oller had applied factor analysis incorrectly thus drawing the wrong conclusions (Bachman, as cited in Chen, 2011).
example, into the factor structure of language test scores: Sawaki, Stricker, and Oranje (2009), for
instance, reported findings that support one higher order factor and four first-order factors for the
four subskills tested in the TOEFL iBT test. Similarly, In’nami and Koizumi (2012) reported two
distinct factors for reading and listening in the revised TOEIC test, which are highly correlated.
These findings justify reporting one overall score and subscores for individual skills. For a more
detailed overview for research into the factor structure of language proficiency based on test data,
see, for example, Stansfield (2008). One has to concede that such studies look into language
proficiency via the “window” of test scores derived from specific test instruments. The scores, at
least in part, are thus influenced by the construct that is reflected in the test structure. High-stakes
tests tend to assess different subskills with distinct test batteries, unlike the integrative tests used
for research purposes in the 1970s. At least in part, the research findings seem to be influenced
by the instruments we use to look at language as well as by the methods of analysis. Hence, the
aforementioned results of recent studies into the factor structure of language test scores come as
no big surprise.

Another relevant line of research is to be found in diagnosing language proficiency to report
strengths and weaknesses for individual learners. Here, a look at detailed subskills is necessary
to report fine-grained and meaningful feedback. Although this is still an underresearched area
and diagnostic language tests are a scarcity (e.g., Alderson, 2005), there are some promising
projects under way. Alderson and Huhta (2011), for example, described three projects investi-
gating the processes taking place in reading comprehension. Two projects look into PISA test
items (native-language [L1] reading comprehension) and reading items from the Pearson Test of
English (second-language [L2] reading comprehension), using the same research questions and
methodology to examine whether L1 reading comprehension models can be applied to L2 reading
tests. The third project, DIALUKI, examines reading processes longitudinally across L1, second,
and foreign languages in three age groups in order to determine reading processes and causes
of reading problems. These projects help shedding light on the nature and conceptualisation of
reading comprehension as one subdimension of general language proficiency. Within this line
of research, cognitive diagnosis is another promising approach to examine to what extent dif-
terent reading skills are identifiable, and in what ways they contribute to discrimination. Here,
employing Q-matrices has demonstrated the potential of gaining insight into the nature and con-
ceptualisation of proficiency underpinned by statistical analyses (Jang, 2009; Lee & Sawaki,
2009; Sawaki, Kim, & Gentile, 2009). Q-matrices are item-incident matrices that aim at defining
the relationship between test items and construct attributes. The attributes are based on cog-
nitive content analysis of test items and on student-reported cognitive processes. Psychometric
modelling of items, attributes, and learner performance helps determining the contribution of
attributes to item discrimination. One of the main purposes of cognitive diagnosis is to “provide
test score users with fine-grained information about learner test performance, so that individual
learners’ strengths and weaknesses can be identified”; therefore, “performance feedback is pro-
vided at a level that is more detailed than the level at which scores are typically reported” (Sawaki
et al., 2009, p. 190). Hence, this line of research opens a promising view on how certain subskills
of language proficiency are conceptualised in language tests.

It is worthwhile to review research done on the nature of language proficiency outside pure
testing contexts. One influential model in the educational school context has been Cummins’s
(1979, 1984, 2008) notion of Basic Interpersonal Communication Skills (BICS) versus Cognitive
Academic Language Proficiency (CALP). BICS denotes conversational fluency whereas CALP
encompasses academic and literacy concepts. This notion has been developed with a view to “timelines and challenges that second language learners encounter as they attempt to catch up to their peers in academic aspects of the school language” (Cummins, 2008, p. 71). Hulstijn’s recent research on language proficiency works along similar lines: He distinguishes between Basic versus Higher Language Cognition (BLC vs. HLC) and core versus peripheral components of language proficiency (Hulstijn, 2010, 2011a, 2011b). However, unlike BICS and CALP, his conceptualisation is based on comparisons between L1 speakers and L2 learners. Hulstijn defines BLC as oral language processing encompassing unconscious knowledge about phonetics, prosody, phonology, morphology and syntax, explicit knowledge in the lexical domain, high-frequency words and structures occurring in communicative situations, and the automaticity of processing. HLC for Hulstijn encompasses less frequent vocabulary, less common structures, extended utterances, and written language; the distinction between BLC and HLC vocabulary and structures is not clear-cut but rather to be seen “in terms of prototypicality” (Hulstijn, 2011b, p. 232). Hulstijn claimed that although all L1 speakers share mastery of BLC, it is an open question to what extent L2 learners may achieve this. It is the area of HLC where L1 speakers show variation; L2 learners may acquire a level of HLC that some L1 speakers may not achieve, as HLC is influenced by educational, intellectual, and motivational factors and personal interests. Complementary to the notion of BLC and HLC, Hulstijn (2011b) proposed a complex model of language proficiency with a core of “linguistic knowledge and the speed with which this knowledge can be processed” (p. 238), with nonlinguistic components at the periphery (strategic and metalinguistic competences).

Current research findings thus indicate that language proficiency is regarded these days as a far more complex concept than the somewhat “simple” notion of a global or divisible nature. It seems best represented by a highly complex, multilayered multicomponent conceptualisation that as yet has not been fully explored and researched in its different contexts such as educational, professional, or assessment settings. In the field of testing, a number of influential proficiency models have been developed since the 1970s (e.g., Bachman, 1990; Bachman & Palmer, 1996; Canale & Swain, 1980, 1981; Hymes, 1972); for in-depth reviews of these models, their strengths and their weaknesses, for instance, with a view to their (lack of) empirical foundations, see, for example, Purpura (2008) or Hulstijn (2011b).

So far, the multilayered componential nature of language proficiency has been discussed, looking—so to speak—at the “horizontal dimension” of proficiency and how it can be divided into more specific subcomponents, such as the four skills or cognitive subcomponents in one of the skills. There is, however, another dimension to take into consideration, that is, the “vertical dimension” of ascending levels of proficiency, which are usually employed for reporting purposes in language tests. Here, the relationship between ascending levels of proficiency in a test situation and a longitudinal view on the development of language proficiency is of interest. Hulstijn

Of interest, there are some tests where aspects of the BLC such as the speed of processing or linguistics elements are operationalized in the speaking part (e.g. Pearson PTE Academic: http://pearsonpte.com/pteacademic/; Versant: http://www.versanttest.co.uk/).

One could conceptualise a third dimension, that is, the yardstick or benchmark against which any language proficiency is assessed. The issue of what standardised or codified language could be used as such a yardstick would be worth further exploration; however, it was not discussed at the LTF and would go beyond the scope of this article. (I am grateful to an anonymous reviewer for this suggestion.)
(2011b) suggested that such levels of proficiency make sense in educational contexts for practical purposes, but they “have no function in a theory aiming at explaining individual differences in the attainment of a first or second language” (p. 242). I agree with him that it is crucial to keep apart levels of proficiency, as assessed in a test with a momentary view on a learner’s proficiency, from levels of attainment, with their focus on the developing, dynamic nature of language learning. This important distinction, however, is not always taken into account in test contexts. This is specifically true for one of the most influential tools referred to across various assessment contexts in Europe and beyond, the CEFR (Council of Europe, 2001). Although the CEFR claims to cover both aspects of proficiency and development in its six main ascending levels of proficiency, it fails to do so consistently (e.g., Alderson et al., 2006; Harsch, 2007; Hulstijn, 2011b; Norris, 2005; Weir, 2005). A number of researchers (e.g., Cumming, 2009; Fulcher, 2004a, 2004b; Hulstijn, 2007; Spolsky, 2008; Weir, 2005) have expressed concerns regarding the foundation of the CEFR scale system. Spolsky (2008, p. 300), for instance, criticised the CEFR as “arbitrary” standard to produce uniformity, whereas Cumming (2009, p. 92) pointed out the dilemma of the imprecision of standards such as the CEFR “in view of the complexity of languages and human behaviour” and their foundation on professional consensus rather than empirical evidence. The “shaky ground beneath the CEFR” (Hulstijn, 2007) becomes particularly significant at the higher levels (B2 and upward). Hulstijn (2011b) argued that the tasks stated at these levels demand intellectual and academic skills, which can only be attained by higher levels of education. These concerns raise doubts about the suitability and adequacy of the CEFR descriptors at the higher levels, having impact for example on high-stakes tests and their demands on test takers who come with a variety of educational and intellectual profiles (Green, 2012). The findings also raise concern about the empirical underpinning and applicability of the CEFR levels. Nevertheless, it has to be acknowledged that the CEFR scales are one of the best researched scales (Hulstijn et al., 2010, p. 15) with a view to their construction from a set of more than 40 existing scales. Teacher judgments were used to Rasch scale and calibrate the descriptors on the CEFR scales (North, 2002; North & Schneider, 1998). The empirical calibration of these scales, however, does not equal an empirical underpinning of the descriptor content or the levels, which could, for instance, be done by analysing actual learner performance.

To address some of the concerns around the CEFR, the SLATE group (Second Language Acquisition and Testing in Europe, see www.state.eu.org) recently formed a network of researchers across Europe who are interested in understanding language development in relation to the CEFR levels (Bartning, Martin, & Vedder, 2010; Roberts, Pallotti, & Bettoni, 2011). Among their aims (Hulstijn et al., 2010) are the analysis of linguistic features of learner performance against the different CEFR levels (i.e., analysis of linguistic profiles for the productive skills and linguistic features for the receptive skills), the examination of differences in profiles across different target languages and different native speaker communities, or the investigation of which linguistic features could be used for diagnosing learner proficiency (here this line of research feeds into the aforementioned underresearched area of diagnostic tests). One overarching aim is to suggest improvements for the CEFR based on the research outcomes of different SLATE projects. Some promising research projects are currently under way (see Bartning et al., 2010, for a detailed account): The Cefling project in Finland examines the development of linguistic features to underpin the CEFR levels, using corpora in Finnish and English; the Communicative Adequacy and Linguistic Complexity (CALC) study looks into writing performance at certain CEFR levels to investigate aspects of communicative adequacy and linguistic complexity; another
Another approach to analysing the CEFR levels can be found in the English Profile programme, a collaborative research programme encompassing a number of projects investigating the Cambridge Learner Corpus and collecting learner language to compile the Cambridge English Profile Corpus complementing the Cambridge Learner Corpus. The programme is built around three major research strands: corpus linguistics, pedagogy, and assessment. Its aim is to create a “profile” or set of Reference Level Descriptions for English linked to the CEFR, describing grammar, vocabulary, and linguistic functions (for further details on research, resources, and events, see http://www.englishprofile.org).

Given the impact and momentum the CEFR is gaining all over the world, and the extent to which the CEFR was discussed at the LTF, the next two sections take a closer look at the CEFR in language testing contexts. The following section explores test scores and the purposes for which they are reported with reference to the CEFR, whereas issues of test alignment to the CEFR are discussed afterward.

TEST PURPOSES, TEST RESULTS, AND THE CEFR

At the LTF 2010 panel discussion it was pointed out that the question of divisibility of language proficiency in a test context would also depend on the level of abstraction that the test developers and the users are interested in. Score reporting and dimensionality can be regarded as two distinct aspects: The former is perhaps more influenced by social and economic demands, whereas the latter is based on empirical evidence. Nevertheless, the two aspects are closely intertwined: the level of abstraction operationalized in the test instrument (i.e., the fine-graininess of test items) ideally matches the test purpose in terms of the level of abstractness on which test scores are reported. For example, test results for international English proficiency tests are usually given in terms of an overall level, complemented by scores on separate skills (sometimes even on separate areas of linguistic knowledge), which are tested in different sections of the test, thus at that level reflecting a divisible view of proficiency. Ultimately, the way test results are reported depends on the users’ demands and the test purpose, which in turn influences how language proficiency is conceptualised. For instance, in certain contexts a complex diagnostic profile may be what is required, with a multidivisible construct of language proficiency underlying the diagnostic test—a characteristic that interestingly enough was discussed by Morrow (1981) more than 30 years ago. Often however, a simple single score for straightforward interpretation is demanded, for example, by administrators or university admissions.

In the LTF discussion it was acknowledged that although some users such as administrators may need a single score, other users like language teachers need more detailed information. The question was raised whether a single exam developer could serve both purposes with the same test, the crucial issue lying in adequate validation and justification of using one test for different purposes (e.g., Fulcher & Davidson, 2009, on validation issues when a test is used for purposes different from its original design; or Spolsky, 2008). Spolsky (2008) argued that reporting one global score is problematic, as for him language proficiency cannot be expressed in one single unidimensional measure: “As we learnt the complexity and the enormous variation of human communicative skills and knowledge, we ignored the challenge to produce interpretable but rich
profiles and agreed to continue to build unidimensional scales” (p. 302). He warned that relying on one single score for high-stakes decisions is dangerous due to the possibility of misclassification and a lack of predictive validity of one single measure. There is a tendency in Europe for expressing single grades in simplistic notions in relation to one overall CEFR level, such as a “B1-person” or a “B1-test.” This bears the danger of simplistic labelling, thus ignoring the complexity of learners and their emergent proficiencies (e.g., Cumming, 2009; Spolsky, 2008). The discussants at the LTF agreed that one of the major challenges test developers and testing researchers face is how to encourage test users and stakeholders to move away from a single grade. One possible answer to this challenge that was proposed time and again during the discussion is the development of assessment literacy among stakeholders (see following the section on assessment literacy).

It also emerged during the LTF discussion that if one chooses to report scores on different levels of abstraction, one would need to apply an adequate multidimensional psychometric model, which can account for uneven profiles in the subskills while also provide a reliable and valid overall score, typically in Europe reported in terms of CEFR levels. There is some promising research done in the field of multidimensional IRT scaling, which can address multidimensionality in a variety of ways (e.g., Ackermann, 1994; Hartig & Höhler, 2009; Osteen, 2010). Such models can, for example, be used to understand what individual items and what the test as a whole are measuring, and they can be used to empirically validate test specifications (Ackermann, 1994). Recently developed models are able to “provide individual ability profiles as test results rather than single test scores” (Hartig & Höhler, 2009, p. 58). Of interest, the question at which level to differentiate which dimension is regarded as a nontrivial one, and Hartig and Höhler, for instance, recommended that “the hierarchical level of the dimensions that are to be measured should primarily depend on the intended use of the results” (p. 60). Although Multidimensional Item-Response Theory (MIRT) models can address multidimensionality, they cannot themselves produce a global score. To report one overall score based on all items, Hartig (personal communication, May 2011) suggested subjecting all items to a unidimensional IRT model.

The LTF discussion then turned to the question of what purposes and audiences a proficiency scale system such as the CEFR can serve. Despite expanding research and heated debates (for a public debate, see, e.g., Fulcher, 2004b; North, 2004), this question remains an open one. A useful distinction of different scale purposes was proposed by Alderson (1991), extended by Pollitt and Murray (1996), namely, the distinction between constructor-, assessor-, user-, and diagnosis-oriented scales. This differentiation can help clarify for both developers and users the scope of the different scale types. Research on the CEFR scales includes evaluations of their intended purposes (e.g., North, 2000, 2007; Schneider & North, 2000), analyses of their applicability to test development contexts (e.g., Alderson et al., 2006, on an analysis with a view to assessing receptive skills; Harsch, 2007, on an analysis addressing writing skills; or Harsch & Martin, 2012, on adapting CEFR scales for rating purposes), as well as reports on their usability in actual test development projects (e.g., the case studies in Alderson, 2002; the collection of reports in

---

6 One could argue for a possible interpretation that the CEFR is a band scheme that does not in and of itself make a claim about dimensionality. However, the background of the CEFR scale system (e.g., the model of communicative competence by Canale & Swain, 1980, and others upon which the scale system is based) and the hierarchical nature of the CEFR scale system strongly imply a multifaceted concept of proficiency (see also Hulstijn et al., 2010). The CEFR scale system offers different levels of abstraction, from the very general overall scale, to overall scales per skill, down to scales describing specific subsaspects of different skills and linguistic components.
Morrow, 2004; or Rupp, Vock, Harsch, & Köller, 2008, on a project evaluating performance standards in line with the CEFR). The literature reports a number of challenges when using the CEFR in test development contexts, implying limitations of the CEFR’s usefulness in analysing features like task difficulty, text types, subskills, or grammatical and vocabulary features. The reported limitations are attributed to issues such as the aforementioned lack of empirical foundation or the insufficiency of the CEFR descriptor system in terms of consistency and terminology (for a summary, see, e.g., Weir, 2005). It is to be anticipated that research on the CEFR foundations and on applying the CEFR in testing contexts will expand over the coming years, hopefully contributing toward a revised CEFR proficiency system.

**TEST ALIGNMENT TO THE CEFR, TEST USAGE AND VALIDATION, TEST EQUIVALENCE**

Although the aforementioned limitations and challenges of using the CEFR in testing contexts are acknowledged in the field of testing research, the CEFR, nevertheless, is employed in high-stakes contexts such as university admissions or immigration decisions. Increasingly, it is used as a means to link examinations to the CEFR and to report test scores in terms of CEFR levels, whether single scores or detailed score-profile reports.

With regards to aligning tests to the CEFR, the LTF 2010 discussion brought to light a shift from the traditional question “Is my B1 your B1?” toward the need for showing “What is your B1 like?” During the LTF discussion, the oversimplification of the first question was acknowledged, with a shift in focus on clarifying and specifying the nature of the relation between a given test and the CEFR, accepting that there are multiple ways of operationalizing “B1” in different contexts and for different purposes. The nature of the relation between an exam and the CEFR could be examined with regard to relevant “comparability” or “alignment” criteria. The Council of Europe has long understood the demands in this area and commissioned the Manual for Relating Exams to the CEFR (Council of Europe, 2003, 2009b; Figueras, North, Takala, Verhelst, & Van Avermaet, 2005) to outline procedures and requirements for alignment. This is now complemented by a Reference Supplement addressing technical issues (Council of Europe, 2009a). The Manual (Council of Europe, 2009b) offers extensive grids to specify and characterise exams on an overall exam level and for separate skills. These grids not only serve as the basis for formal test linkage procedures but also help communicating what a given exam is covering in terms of constructs, content, and test design characteristics, thus providing the aforementioned comparability criteria.

The shift in focus that emerged from the LTF discussion emphasises the importance of test providers’ responsibilities of demonstrating the “B1-ness” of tests that claim to be linked to the CEFR levels in order to fulfil the requirements of backing up claims with evidence as part of a test’s validation and validity argumentation as advocated, for example, by Kane (2001, 2010, 2012) or Messick (1989). The resulting transparency could help investigating the important issue of the comparability of scores reported by different tests that claim to be linked to the CEFR. Here, the CEFR can serve as tertium comparationis, without supporting any claims of “comparability” or “equivalence” in terms of constructs or content being covered. The authors of the Manual (Council of Europe, 2009b, p. 4) point out that different exams (claiming to have been) linked
to a certain CEFR level do not necessarily have to be testing the same constructs, nor would this imply the tests are in some way equivalent or comparable. The LTF discussion made it clear that it is important that examination bodies develop and set out their argumentation to justify what they are claiming, and show how the evidence, counterevidence and counterarguments support or challenge their claims. Whether this is one of the CEFR’s intended purposes or not seems irrelevant relative to what is at stake here, given the fact that the CEFR is actually used for this purpose. The plenary discussion at the Language Testing Forum 2011 took this crucial point up again, showing that it is an ongoing and yet unresolved issue.7

Research and reports on aligning tests and exams to the CEFR, both in qualitative and psychometric ways, are thriving. The literature shows an increasing number of research studies on alignment projects across a range of contexts, settings, and foreign languages, also beyond Europe (e.g., Bechger, Kuijper, & Maris, 2009; Figueras & Noijons, 2009; Glaboniat, Perlmann-Balme, & Studer, 2013; Harsch, Pant, & Köller, 2010; Martyniuk, 2010; Noijons & Kuijper, 2006; Papageorgiou, 2007, 2010; Tannenbaum & Wylie, 2008; Taylor & Jones, 2006; Wu & Wu, 2010). The vast majority of these projects makes use of the procedures outlined in the Manual (Council of Europe, 2009b) and uses the specification grids offered there to describe what their tests are measuring in which ways.

Alignment to a framework such as the CEFR brings the advantage of adding meaning to test scores (Kane, 2012, p. 8) as the CEFR can-do statements can be used to report what learners can do at a certain proficiency level. This, in turn, “adds an extra layer of interpretation” (Kane, 2010, abstract) to the score interpretation process. If test scores are to be interpreted in terms of CEFR levels and their can-do statements, evidence has to be collected to show in what ways and to what extent aspects such as the task demands, the constructs of the test items, the cognitive processes elicited or the elicited performances match and operationalize the statements defining the CEFR levels. Figueras and Noijons (2009), for example, compiled a number of research studies that reported issues with the alignment procedure, such as judges having difficulty rating performances against CEFR levels (Maris, Noijons, & Reichard, 2009), difficulties for setting standards for grammar and vocabulary items (Downey & Kollias, 2009), or not enough guidance on item characteristics for specific levels (Moe, 2009). At the LTF discussion, the concern was raised of a need for psychometric expertise for choosing and applying appropriate standard setting methods. Further measurement issues that need to be taken into account are the use of sophisticated procedures and models to analyse the relationship between test scores, characteristics (or specifications) of tests, items or performances and the targeted CEFR level (e.g., Kane, 2001), as well as measurement models to evaluate the quality of judgmental linkage procedures; for an example of such a measurement model implemented for the C-test in TestDaF, see Eckes (2012). Moreover, when trying to interpret test scores with reference to the CEFR, issues of cognitive validity (Borsboom, Mellenbergh, & Van Heerden, 2004; Weir, 2005) need to be taken into account. Weir (2005), for example, raised the issues that the CEFR does not allow to coherently apply contextual variables and performance conditions across varying levels of proficiency; that cognitive processing is only marginally accounted for; and that there is no link of activities to expected outcome qualities, in addition to the inconsistent descriptor wording. These issues make mapping tests against the CEFR a challenge.

---

7The summary of the discussion of language tests for immigration purposes is available at http://www2.warwick.ac.uk/fac/soc/al/research/conferences/ltf2011/language_testing_forum_2011_panel_discussion.pdf.
Furthermore, some researchers see an issue with the validity of the framework itself. McNamara, for instance, regarded the CEFR scale system as being the “product of political forces—not academic argument” or empirical validation (McNamara, 2006, p. 37). Chapelle (2012), taking a similar line, called alignments to frameworks such as the CEFR “controversial because they attempt to connect social and political meanings associated with frameworks with the scientific procedures used to understand score meaning” (p. 25). One way forward could be to make use of “good principles” in test development and validation, as for example laid out in Bachman and Palmer’s (2010) Assessment Use Argument, which provides “an explicit rationale for the links between assessment performance and assessment use” (Bachman, as cited in Chen, 2011, p. 284). Justifying assessment use right from the beginning of a test development project up to the alignment process and beyond into the actual usage of the test can help communicate transparently the validated areas of usage for stakeholders and future test users. Although developing such assessment use arguments seems straightforward for new test developments, the issue remains for existing tests, particularly if they are used for purposes other than the original ones. Here, the suggestions made in Fulcher and Davidson (2009) with regard to test retrofit are particularly relevant: Each change of purpose has to be clearly articulated in order to construct a new validity argument for the new test purpose. A failure to do so will result in limited validity.

Especially in contexts where standards are in place across a large region and where there are competing tests from different providers “freely” available (albeit on a commercial basis), the question of test comparability and test equivalence becomes paramount (e.g., Cumming, 2009). In recent years, in Europe and beyond, the CEFR has taken on the role of a standards-based comparison criterion for high-stakes purposes such as university admissions or migration decisions, leading to the aforementioned simplistic notion of a “B1-test.” Some test users (wrongly) assume that tests claiming to be linked to a certain CEFR-level have to be testing “the same” skills or knowledge elements, which is a notion not supported by the authors of the Manual, as previously discussed. Nevertheless, users demand simple and easily interpretable conversion tables between different high-stakes tests (ideally for just one single score with the previously discussed implications of simplified concepts such as a “B1-learner”). Such conversion tables, however, would not only presuppose that the tests in question have been adequately developed and validated for the same specific use, and aligned to the CEFR adhering to rigorous standards, but also require comparability studies between the different tests. Reports on such studies are scarce; they include the TOEFL-Cambridge Comparability Study (Bachman, Davidson, Ryan, & Choi, 1995), a comparability study done during the pilot phase of the Pearson PTE Academic with IELTS and TOEFL iBT (Pearson, 2009, pp. 44–50), or a recently published study in the

---

8 Although the CEFR scales have been empirically calibrated using teacher judgements (North, 2002), this does not amount to a validation of the scales for specific purposes, such as assessor-oriented or constructor-oriented purposes. Similarly, the statements in CEFR Chapter 9 (Council of Europe, 2001, pp. 180ff) that the proficiency scales might be useful for such purposes would need to be backed up by empirical validation.

9 An example for a non-European context can be found in Hong Kong, where the Workplace English Campaign was set up in 1999, supported by the Standing Committee on Language Education and Research; one of its aims was to “benchmark English standards for different occupational needs” (Qian, 2008, p. 102). Workplace English Campaign has developed five proficiency levels, which are claimed to be aligned to the ALTE and CEFR levels, and Workplace English Campaign recognises a number of international English language tests for employer information and to identify training needs. Alignment tables can be accessed at http://www.english.gov.hk/eng/html/wec_hkweb_alignment_table.htm. Knight (2001) reports a case study on the benchmarking.
Taiwanese context on comparing GEPT with IELTS and real-life academic performance (Weir, Chan, & Nakatsuhrara, 2013). Results are not conclusive at present, and the demand for such comparability studies is clearly given (see, e.g., the outcome from the LTF 2011 panel discussion). Responsibilities for such studies should be shared in a collaborative way between researchers, developers, and end users (e.g., Stansfield, 2008).

The overarching ethical issue of responsibilities for developing test use justifications, for rigorous test development and validation procedures, for a transparent documentation of alignment and comparability studies, and for using tests within their validated realm is best addressed by sharing and distributing responsibilities between developers, researchers, users, and stakeholders. Some areas lie clearly in the domain of a specific group (e.g., test validation should be the responsibility of researchers), but Davies (2012), for example, stated that the responsibility for using tests within their validated purposes cannot “be laid at the door of the researcher” (p. 41) but lies with the test user. It has to be acknowledged, however, that not all test users share the necessary knowledge of assessment issues, their main interest often lying in the information about what score they should use for a specific purpose (e.g., admissions). Hence, this lack of knowledge could become an issue that needs to be addressed when arguing for shared responsibilities.

I would like to agree with Stansfield that our profession of testing researchers can make an important contribution here: We have “to accept social and political responsibilities” (Stansfield, 2008, p. 323) for educating stakeholders and developing assessment literacy.

ASSESSMENT LITERACY

As became apparent during the LTF 2010 discussion (and in fact during the discussion the following year), assessment literacy is one of the answers to the existing challenges just outlined. Stansfield (2008) reminded us of our responsibility as a profession to pass on our professional knowledge and skills to all stakeholders involved. In recent years, there have been a number of surveys and needs analyses looking into the current state of assessment literacy among teachers and other stakeholders (e.g., Hasselgreen, Carlsen, & Helness, 2004; Huhta, Hirvelä, & Banerjee, 2005; Vogt, 2009), revealing a general need for developing such literacy, as not many stakeholders reported familiarity or formal training. There are also accounts of what aspects and components assessment literacy, and more specifically language assessment literacy, should encompass (e.g., Brindley, 2001; Brookhart, 2011; Inbar-Lourie, 2008; Leung, 2004; Malone, 2008; Taylor, 2009). Brindley (2001), for instance, described core and complementary modules in a course of assessment literacy. With regard to such courses preparing foreign language teachers in China, a study by Jin (2010) revealed adequate coverage of essential testing aspects; however, it emerged that measurement issues and assessment practice in the classroom were not given enough attention. There are also two surveys and analyses of existing assessment courses by Bailey and Brown (1996) and Brown and Bailey (2008), comparing results across one decade. The researchers conclude that although there is a stable core basis of assessment-related knowledge, there have been considerable changes over the two points of research. Nevertheless, Brindley’s (2001) work and the studies by Bailey and Brown could offer a suitable starting point for describing core knowledge and skills valid across a variety of different local contexts and settings.

Existing “guidelines of good practice” in (language) assessment are a further source for promulgating assessment literacy, such as the ALTE Principles of Good Practice (Association of
Language Testers in Europe, 2001), the EALTA Guidelines (European Association for Language Testing and Assessment, 2006), the ILTA Guidelines for Practice (International Language Testing Association, 2007), or the JLTA Code of Good Testing Practice (Japan Language Testing Association, n.d.). It is a promising development to see test providers developing their internal quality standards (e.g., Educational Testing Service, 2002), or documenting their development and alignment procedures with reference to such standards (e.g., De Jong & Zheng, 2011).

With regards to educational and teacher competencies, there are, for example, the Standards for Teacher Competence in Educational Assessment of Students (American Federation of Teachers, National Council on Measurement in Education, & National Education Association, 1990). There are some tools to evaluate such competencies, such as the Classroom Assessment Literacy Inventory (Mertler, n.d.) targeted at the American context, or the general Teaching Knowledge Test (Cambridge ESOL, n.d.), but to my knowledge no tool for evaluating a common core of language assessment literacy has yet been developed. To promulgate assessment literacy within the teaching profession, the most promising way lies in incorporating assessment courses into pre- and in-service teacher training, making assessment an integral part of teacher education, and enabling access to the materials listed here. Furthermore, a new and significant perspective is taken by Watanabe (2011), who reported a course in assessment literacy for test takers as “the most important stakeholders” (p. 29).

Last but not least, existing text and research books on language testing and assessment (e.g., Bachman & Palmer, 2010; Douglas, 2010; Fulcher, 2010) are another rich source for developing assessment literacy among all participants in the assessment process.

**CONCLUSIONS—A RESEARCH AGENDA FOR THE FUTURE**

As far as research into the conceptualisation and development of language proficiency is concerned, Hulstijn’s (2011b) model proposes an empirically testable set of hypotheses with the following research implications. Further comparison studies are needed between L1 and L2 speakers, controlling age, educational and intellectual background in order to yield results based on L1 and L2 groups with similar background profiles. Cummins (2008) proposed to further explore the relationship between language proficiency and academic development, which ties in with exploring Hulstijn’s suggested differentiation between basic and higher language cognition. Hulstijn (2011b) furthermore suggested research into “uneven profiles” (p. 243) to explore the extent of unevenness and the compensatory nature of different components especially at higher levels of proficiency. With a view to lower levels of proficiency, further research into minimally necessary linguistic competencies is needed in order to find out what role, for example, vocabulary plays at these levels. There is some research on vocabulary under way in the context of the SLATE network (e.g., Milton, 2010). Research is needed both for productive and receptive skills, capturing the developmental nature of evolving competencies in longitudinal studies as well as covering different levels of proficiency. The use of corpora for the latter area is highly promising and is currently being employed by some of the SLATE researchers: Corpora of authentic learner language (collected outside pure testing contexts, as is done in the Cefling project) could help shed light on the structure and nature of productive language proficiency, whereas general language corpora could enhance our knowledge of the characteristics of tasks targeting different levels of receptive proficiency. Current and ongoing research projects in these areas are reported in Bartning et al. (2010) and Alderson and Huhta (2011).
When it comes to assessing language proficiency, innovative ways of assessing integrated tasks could be explored, along with new measurement models that can account for the intertwined and complex nature of language proficiency. As previously indicated, if test users continue to demand single score reports alongside profile scores, adequate psychometric models such as the aforementioned MIRT models (Ackermann, 1994; Hartig & Höhler, 2009; Osteen, 2010) need to be employed that can account for uneven profiles in the subskills, yet come to a reliable and valid overall score. With respect to innovative ways of testing, studies into the advantages of such innovations could help decide which way to follow for which contexts: Spolsky (1995, 2008), for example, suggested using a combination of proficiency and aptitude tests for high-stakes decisions, whereas Hulstijn (2011b) proposed testing in distinct stages from an initial test targeting core language elements, which (when taken successfully) is followed by a test targeting peripheral elements in order to save time and resources. Admittedly in the latter case, issues of fairness (a test taker can sit the peripheral test only once the core test has been taken successfully, which could disadvantage test takers with a high ability in peripheral elements) would have to be taken into consideration. Such a novel approach to testing proficiency requires more research in order to understand its validity better.

The important issue of the role of the CEFR in language assessment leads to a range of research desiderata with regard to the instrument itself, hopefully resulting in a revision of the framework with a view to an empirically derived basis (beyond teacher judgments) for the components and levels, and with a view to research-based validation procedures. The CEFR descriptors and their placement at specific levels could be explored in relation to the previously proposed studies into the nature and development of language proficiency and academic competencies; some promising projects are currently being carried out by the SLATE network and the English Profile as detailed here earlier. With a view to researching the use and applicability of the CEFR for different contexts and purposes, more research is needed to evaluate on empirical grounds whether one framework can fit all contexts and purposes, or whether certain adaptations are needed for certain contexts, languages or user groups. Norris (2005), for example, suggested to explore “how—and how effectively—test developers and users are putting the CEFR scales and descriptors into practice in local settings” (p. 405). Here, besides the aforementioned case studies (Alderson, 2002) and research reports on standard setting projects (Figueras & Noijons, 2009), impact studies would be helpful from contexts where the CEFR has been the basis of reform in educational settings. With regard to the applicability of the CEFR for different languages, studies on the comparability between CEFR proficiency levels and proficiency levels formulated for non-European languages (e.g., tonal or logographic languages such as Chinese) are needed. The use of the CEFR in testing contexts could be a further focus, as it is far from ideal for testing purposes (e.g., Alderson et al., 2006). Here, the relationship between specific characteristics (test, task, or item related) and certain CEFR levels could be systematically examined with a view to developing a classification system similar to the Dutch Grid10 (Alderson et al., 2006), but with a link of certain (combinations of) characteristics to certain CEFR levels. One ongoing project examining test characteristics for reading and listening tests with a view to linking them to the CEFR is reported by Harsch and Hartig (2012).

10 The Dutch Grid is a classification system to characterise tests of listening and reading comprehension; the classification categories and options are derived from the CEFR. The Grid aims to facilitate test specification in order to relate the tests to the CEFR. The Grid is accessible online at http://www.lancs.ac.uk/fss/projects/grid.
Another research area is the field of exam linkage to the CEFR and test comparability between competing high-stakes tests. Here, the “big” exam providers are asked to improve transparency on these procedures in order to provide research-led information on the “B1-ness” of their tests. Studies on the impact of certain test use contexts, such as using language tests targeting university admission for immigration purposes, are needed in order to evaluate ethical, social, and political consequences. To derive empirically based conversion tables for test comparison purposes with reference to the CEFR, more cross-test alignment and comparability studies are required. They could also serve to investigate whether test takers have equal chances of a comparable classification by different tests. Although these studies are important to inform stakeholders and test users on empirical grounds, they could potentially mislead to an “overinterpretation” of the comparability: Often, it is only the level of difficulty that is comparable, not necessarily test content or the underlying constructs. Hence, such comparability studies need to be clearly communicated to stakeholders in their potentials and limitations.

This leads us to the last area I would like to propose for further research: assessment literacy. In this field, it would be desirable to compile a list of core knowledge and skill areas upon which assessment courses could be based, taking into consideration the work already accomplished here (e.g., by Brown & Bailey, 2008, or by Brindley, 2001). This core can be complemented by specific areas of knowledge and skills derived from needs analysis for local contexts. Exploring feasible ways to disseminate necessary information and training to certain stakeholder groups is also needed, along with impact studies of such dissemination. Needless to say that assessment literacy should also encompass assessment numeracy in order to understand the contributions and limitations that psychometrics has to offer. It could be worthwhile to develop instruments for evaluating the extent of existing assessment literacy for both the core basis and locally required aspects. It is in the area of enhancing and developing assessment literacy that our profession can make a real impact, contributing to ethical and responsible test use and test interpretation.

REFERENCES


Downey, N., & Kollias, C. (2009). Standard setting for listening, grammar, vocabulary and reading sections of the Advanced Level Certificate in English (ALCE). In N. Figueras & J. Noijons (Eds.), Linking to the CEFR levels: Research perspectives (pp. 125–130). Arnem, the Netherlands: CITO, CoE, EALTA.


Maris, G., Noijons, J., & Reichard, E. (2009). Benchmarking of videotaped oral performances in terms of the CEFR. In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR levels: Research perspectives* (pp. 81–85). Arnhem, the Netherlands: CITO, CoE, EALTA.


