剑桥语言测试研究丛书

Aligning Tests with the CEFR

Reflections on Using the Council of Europe's Draft Manual

以欧洲语言共同框架为基准的语言测试：

关于欧洲委员会工作草案施行的思考

Waldemar Martyniuk 编

上海外语教育出版社
外教社 SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS
www.sflep.com

剑桥语言测试研究丛书

Aligning Tests with the CEFR

Reflections on Using the Council of Europe's Draft Manual

以欧洲语言共同框架为基准的语言测试：

关于欧洲委员会工作草案施行的思考

Waldemar Martyniuk 编

上海外语教育出版社
外教社 SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS
www.sflep.com

图书在版编目（CIP）数据

以欧洲语言共同框架为基准的语言测试：关于欧洲委员会工作草案施行的思考/瓦尔德马·马丁纽克编.
—上海：上海外语教育出版社，2018
（剑桥语言测试研究丛书）
ISBN 978-7-5446-5413-5

I. ①与… II. ①瓦… III. ①英语—测试—研究 IV. ①H319

中国版本图书馆CIP数据核字（2018）第109320号

图字：09-2017-927号

出版发行：上海外语教育出版社
（上海外国语大学内） 邮编：200083
电　　话：021-65425300（总机）
电子邮箱：bookinfo@sflep.com.cn
网　　址：http://www.sflep.com
责任编辑：张亚东

印　　刷：启东市人民印刷有限公司
开　　本：635×965　1/16　印张 18.75　字数 385千字
版　　次：2018 年 12月第 1版　2018 年 12月第 1次印刷
印　　数：2 100 册

书　　号：ISBN 978-7-5446-5413-5 / G
定　　价：59.00 元

本版图书如有印装质量问题，可向本社调换
质量服务热线：4008-213-263　电子邮箱：editorial@sflep.com

出版说明

语言测试作为语言教学的重要组成部分，一直是社会关注的焦点。近年来，我国学术界对语言测试研究的关注也在逐渐升温。国际上，学者们对语言测试的研究已经达到较高的水平，无论是从理论框架的形成到复杂数学模型的建构及其应用方面都取得了实质性的进展，而且在对有国际影响力的外语考试（如雅思、托福等）的信度、效度的研究方面也取得了不俗的成果，出版了不少优秀的著作。

英国剑桥大学出版社出版的“Studies in Language Testing”是其中较有影响力的系列之一。丛书由剑桥语言测评中心策划推出，从测试领域的各个研究视角入手，展示了该领域研究的设计、数据收集和数据分析等环节的基础理论与实践操作，反映了测试领域的最新研究动态。

外教社从中精选9种，引进出版。这些图书中，有对语言测试听说读写等方面的效度和质性研究，也有对测试反拨效应及其对教学影响的研究，还有对雅思考试和欧洲语言共同参考框架的分析思考。希望丛书的出版可以为我国语言测试研究领域的发展和创新提供具有借鉴意义的启示，打开更广阔的研究思路。

Aligning Tests with the CEFR

Reflections on using the Council of Europe's draft Manual

Edited by
Waldemar Martyniuk

Acknowledgements

I would like to express my thanks to all contributors to the volume for developing and writing up their original presentations given at the Cambridge Colloquium in December 2007 and for their willingness to make subsequent revisions in line with the editorial suggestions.

The volume could not have reached publication without the professional assistance of the editors of the SiLT series, Mike Milanovic and Cyril Weir, as well as Lynda Taylor of Cambridge ESOL acting on their behalf. I am grateful to them for their support throughout the editing process.

I am also grateful to the whole team at the Language Policy Division of the Council of Europe in Strasbourg led by Joe Sheils for enabling me to become actively involved in the work of this great organisation.

Waldemar Martyniuk

The publishers are grateful to the copyright holders for permission to use the copyright material reproduced in this book, and to Trinity College London for the use of data, diagrams and tables previously published in Papageorgiou 2007.

Series Editors' note

This Series Editors' note is longer than usual, given the importance of the impact that the Common European Framework of Reference for Languages (CEFR) is having on language education in Europe. There is growing interest worldwide in establishing comparability between assessment tools and external standards, whether these are technical standards relating to matters of quality assurance, or descriptions of performance levels that offer benchmarks for comparison. Such standards tend to be influential since they provide policy makers with tools that can be used for gathering baseline data, for benchmarking and for evaluating current practices. External standards are of particular benefit to governments which have educational or test reform initiatives. Given the scope for abuse or at least misuse of the CEFR in this context we feel that an in-depth consideration of the issues involved is an important preface to this volume.

The CEFR levels referred to in this volume are not 'standards' in the true sense of the concept; they form a useful framework of reference and offer a helpful metadiscourse. The word 'standard' is not used in the title of the CEFR, and the introductory notes heavily downplay the notion that the CEFR offers 'standards', though the message is mixed: for example, on page xiii, towards the end of the fourth paragraph (Council of Europe 2001) we read:

> It is already clear however, that a set of common reference levels as a calibrating instrument is particularly welcomed by practitioners of all kinds who, as in many other fields find it advantageous to work with stable, accepted standards of measurement and format.

The CEFR is a widely used, common framework of reference based on six broad reference levels and an 'action-oriented' approach to language teaching and learning. Within a relatively short period of time it has become highly influential in Europe and beyond as a helpful way of articulating objectives for language teaching and learning. The CEFR has certainly helped to raise awareness of language issues and has provided a useful focus for researchers, policy makers, assessment providers, and teachers.

However there is some concern that the CEFR has been adopted or interpreted as a fixed standard or set of standards, even though it perhaps was not originally designed as such. Over time, the pressure has grown, often from government, for test providers and examination boards to link their

examinations to a particular external standard, namely the CEFR. The case of Taiwan of China (see Wu and Wu in this volume) is a good example of this.

In response to perceptions and expectations that the CEFR could and should offer a set of stable and acceptable standards for testing and assessment purposes, the Council of Europe set about providing a 'toolkit' of resources to inform and facilitate the process of aligning tests to the framework. This initially included a draft pilot Manual for *Relating language examinations to the CEFR* and a technical reference supplement to this (Council of Europe 2003, 2004), with a later revised version of the Manual (2009). The Council also provided forums where practitioners could share their reflections on the use of the draft Manual and their experience in using the different linking stages as suggested within it. Examples of such forums include a seminar entitled 'Reflections on the use of the Draft Manual for Relating Language Examinations to the CEFR: Insights from Case Studies, Pilots and other projects' held in Cambridge in December 2007. This *Studies in Language Testing* (SiLT) volume contains many of the papers that were first delivered at that meeting. It provides a number of perspectives on the process and outcomes of attempts to align examinations to the Common European Framework of Reference (CEFR) using the Manual provided by the Council of Europe. Waldemar Martyniuk, the Editor of this volume, outlines the content and focus of these papers in his editorial introduction below.

The Association of Language Testers in Europe (ALTE) and Cambridge ESOL in its own right, and also as a founder member, have encouraged the development of the 'toolkit' to allow users to make better use of the CEFR for their own purposes, and have overseen or been directly involved with a number of initiatives to assemble the necessary resources for the toolkit. These include:

- Co-ordinating the development of a *Users' Guide for Examiners* (1996) – now under revision by ALTE again as a manual for test development and examining (2010).
- Developing the EAQUALS/ALTE European Language Portfolio (ELP), both in hard copy and electronic forms (from 2000).
- Providing support for the authoring and piloting of the draft Manual for relating examinations (since 2002/03).
- Contributing to benchmarking materials and examples of test items and tasks to accompany the CEFR (from around 2004).
- Developing content analysis grids for speaking and writing materials (based on ALTE projects dating back to 1992); and
- Specifically as Cambridge ESOL, playing a co-ordinating role in developing Reference Level Descriptions for English – The English Profile Programme (since 2005). The Profile in English will contribute significantly to the usefulness of the CEFR as a practical tool.

Cambridge ESOL's involvement with the Council of Europe has an even longer history dating back to 1980 when the concept of a multi-level system of Cambridge examinations began to emerge in light of Wilkins' work on proficiency levels (see Wilkins in Trim 1978) and starting with the addition of a Threshold Level test (Preliminary English Test – PET) to the well-established First Certificate in English (FCE) and Certificate of Proficiency in English (CPE) examinations. In 1990 the revised Waystage and Threshold specifications (which had been partly sponsored by Cambridge) formed the basis of the test specifications for the new Key English Test (KET) and updated PET, and further additions and revisions to existing examinations saw the process of convergence taking place to achieve this goal, as noted by North (2008:31–32).

From the early 1990s Cambridge ESOL, both in its own right and also as part of ALTE, worked to develop an empirically derived common scale that allowed for the systematic ordering of its examinations according to level (see the Series Editor's note in *Studies in Language Testing* Volume 1, 1995) as well as the comparison of examinations across languages on the ALTE 5-level system. In the Cambridge context, the empirical underpinning for the system was achieved by introducing an item banking approach which involves assembling a bank of calibrated items – that is, items of known difficulty. Designs employed for collecting response data ensure a link across items at all levels. The Cambridge ESOL Common Scale, a single measurement scale covering all Cambridge ESOL levels, has been constructed with reference to these objective items. The Common Scale thus relates different testing events within a single frame of reference, greatly facilitating the development and consistent application of standards.

Since the inception of the Common Scale many millions of candidates at all proficiency levels have taken the Cambridge examinations and their responses have allowed the scale to be incrementally refined based on analyses of this data within the framework. (See the paper for the Council of Europe by North and Jones (2009) to accompany the revised Manual; also Maris (2009) for discussion of test equating using IRT in the context of standard setting in the collection of papers edited by Figueras and Noijons (2009).) As part of ALTE, Cambridge contributed to the development of the ALTE 5-level system. Underpinning this work is a system of test content analysis and the application of an item banking approach that is applied to the examinations of a number of ALTE members.

It is important, nevertheless, to constantly remind ourselves that the CEFR itself is deliberately *underspecified* and *incomplete* (see Milanovic 2009). It is precisely this feature which makes it an appropriate tool for comparison of practices across many different contexts in Europe and beyond. On the one hand it is useful as a common framework with six broad reference levels, but on the other it is not applicable to all contexts without user intervention in order to *adapt it flexibly* to suit local purposes.

The three main authors of the CEFR, Daniel Coste, Brian North, and John Trim, made this point very clearly in the text itself and they have all repeated it on numerous occasions in subsequent presentations on the framework and its principles. So, for example, in the introductory notes for the user, the following statement is emphatically made: 'We have NOT set out to tell practitioners what to do or how to do it' (Council of Europe 2001:xi). This is reiterated throughout the text by the use of the phrase: 'Users of the framework may wish to consider and where appropriate state . . .' (e.g. page 40).

Subsequent work on the 'toolkit' has also followed this lead. For example, the authors of the Manual for *Relating language examinations to the CEFR* stress this point when they state that the Manual is not the only guide to linking a test to the CEFR and that no institution is obliged to undertake such linking.

More recently, in his plenary paper presented at the Council of Europe Policy Forum on use of the CEFR (Strasbourg 2007), Coste (2007) described how contextual uses which are seen as deliberate interventions in a given environment can take 'various forms, apply on different levels, have different aims, and involve different types of player'. In his view: 'All of these many contextual applications are legitimate and meaningful but, just as the Framework itself offers a range of (as it were) built-in options, so some of the contextual applications exploit it more fully, while others extend or transcend it.'

When considering test alignment questions, this fundamental principle must be borne in mind because there are important implications which follow on from this. For example, it is important to remember that the CEFR is not intended to be used prescriptively and that there can be no single 'best' way to account for the alignment of an examination within its own context and purpose of use. As Jones and Saville (2009:54–55) point out:

> . . . some people speak of applying the CEFR to some context, as a hammer gets applied to a nail. We should speak rather of referring a context to the CEFR. The transitivity is the other way round. The argument for an alignment is to be constructed, the basis of comparison to be established. It is the specific context which determines the final meaning of the claim. By engaging with the process in this way we put the CEFR in its correct place as a point of reference, and also contribute to its future evolution.

A particular concern relates to the status of the 'illustrative scales of descriptors' as they are called, and their recent uses in overly prescriptive ways (i.e. against the intentions of the authors), especially in the context of standard setting. In one of the pre-publication drafts of the framework document entitled *Learning, Teaching, Assessment. A Common European Framework*

of Reference (Strasbourg 1998), these scales were included in the appendix as examples and did not occur in the body of the text. The only scales to be included in the main text were the common reference levels (later to become Tables 1, 2 and 3 in the published version, Council of Europe 2001:24–29).

This original layout of the text in the 1998 draft visibly reinforced the different status and function of the general reference levels and the more specific *illustrative* scales. This approach underlined the very tentative nature of the illustrative scales, many of which were uncalibrated and indeed were underrepresented, particularly at the higher C-levels. Given the vigour with which some people have recently attempted precise alignment using these scales, despite their obvious and clearly stated deficiencies, it is dangerous to give the illustrative scales too much prominence.

In Chapter 8 of the 1998 draft version which was entitled 'Scaling and Levels' the tentative status of the illustrative scales was made clear in the following paragraph (page 131):

> The establishment of a set of common reference points in no way limits how different sectors in different pedagogic cultures may choose to organise or describe their system of levels and modules. It is also to be expected that the precise formulation of the set of common reference points, the wording of the descriptors, will develop over time as the experience of member states and of institutions with related expertise is incorporated into the description (1998:1,310).

Since the publication of the CEFR in its finalised form in 2001, the second point in this paragraph, emphasising the tentative nature of the illustrative scales, has tended to be forgotten or at least downplayed by some users. This may be in part due to the way that the final text was edited. Many of the less well validated illustrative scales remained in the final text, but for pragmatic reasons the authoring group decided to incorporate them into the main text rather than keep them in the appendix. Four appendices were used to illustrate several projects involving the development of scale descriptors; Appendix B (Council of Europe 2001:217) was used to describe the development of the 'illustrative scale descriptors' which was part of the Swiss research project conducted by North (published in 2000 as a book based on his PhD).

But the points made by the authors in 1998 still remain true; in other words, the functional and linguistic scales were there to illustrate the broad nature of the levels rather than to define them precisely. While some of the scales might prove stable across different contexts, there should not be an expectation that they all will. This has important implications for the use of the 'illustrative scales of descriptors' in alignment procedures; for example, given their status, individual scales should only be used with great care in any kind of standard-setting exercise. Indeed it is hard to see how, over and

above a very general approximation to the levels, standard setting using the current scales can be considered an entirely satisfactory procedure.

North himself (2007) notes that the 'fluency' scale was useful in linking the ALTE 'Can Do' project to the framework (based on values from the Swiss project he had conducted) but that other scales were not robustly calibrated, and there were significant gaps at the A1 and C levels (see North's presentation made at the 23rd ALTE conference, Sèvres, April 2007 – available from the ALTE website: www.alte.org).

Somewhere along the way, these very real concerns expressed by a principal author of the scales have been lost or ignored. Indeed, given the origins and status of the scales it is perhaps unfortunate that there has been a somewhat one-sided reading of the text of the CEFR, as noted by Coste (2007), another of its original authors: 'In various settings and on various levels of discourse . . . people who talk about the Framework are actually referring only to its scales of proficiency and their descriptors.' Trim echoes this view in Saville (2005) – *An interview with John Trim at 80*.

In any case, it is important to note that the illustrative scales in the CEFR are precisely that. They are underspecified at the upper levels at least and uncalibrated in many instances. They should be viewed and used with caution, particularly in standard-setting exercises, since they are likely to prove misleading at best and quite damaging at worst.

Embedded procedures rather than one-off exercises

If the CEFR is to have a lasting and positive impact in the context of assessment, then its principles and practices need to be integrated into the routine procedures of assessment providers so that alignment arguments can be built up over time as the professional systems develop to support the claims being made (for examples of how this can be done, see O'Sullivan, and Khalifa, ffrench and Salamoura in this volume). This entails working with the text of the CEFR *as a whole* and *adapting it* where necessary to suit specific contexts and applications. It is unlikely that any single study or report can provide satisfactory evidence of alignment. On the contrary, a single standard-setting exercise should *not* be taken as sufficient evidence of alignment and examination providers should seek to provide multiple sources of evidence accrued over time.

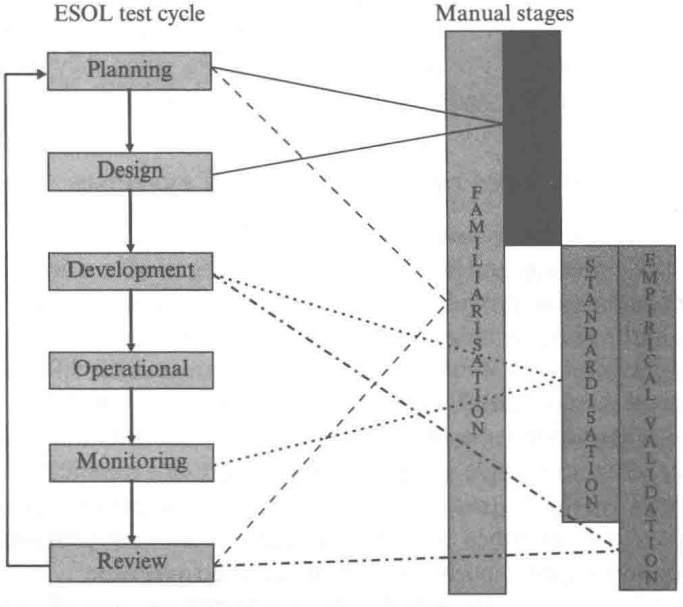
Standard-setting events which are conducted as one-off procedures simply do not provide enough evidence for consistent interpretation of any level system. If necessary, alignment arguments should remain tentative and be modified later in light of additional evidence as and when it becomes available. This should be expected rather than be seen as a problem.

In relation to assessment, therefore, alignment arguments and assessment standards need to be maintained in the long term using a range of techniques and professional processes, including:

- item banking to establish common measurement scales and to allow for both item-based and person-based equating to be employed in test construction and in the monitoring of standards over time
- routine test validation processes to quality assure test outcomes
- iterative cycles of test development and revision.

More specifically, this means that the recommendations found in the Manual on how to use the CEFR and other 'toolkit' resources supplied by the Council of Europe for alignment purposes (e.g. familiarisation activities with stakeholders and standard-setting exercises of different types, whether task- or person-based) need to be *integrated within the standard procedures* of the assessment provider rather than seen as 'one-off events'. Chapter 7 of Khalifa and Weir (2009) in the *Studies in Language Testing* series provides an informative account of how these linking devices are embedded in Cambridge ESOL standard procedures.

Manual procedures and the Cambridge ESOL test cycle



Such an approach is relevant across a broad range of contexts from classroom testing to the work of an examination board like Cambridge ESOL or other members of ALTE which interact with (literally) thousands of stakeholders to develop, administer, mark and validate many different types of examination within a consistent but evolving frame of reference. In 2010, for example, many hundreds of administrations of different language examinations by

Cambridge and other ALTE partners will take place, most of which include the assessment of four skills (including face-to-face Speaking tests). Given the complexity of these operations, the arguments for alignment to external reference points need to be developed on a case-by-case basis and must be one part of the broader validity argument which is needed to support the appropriate uses of each examination.

Finally we would like to return to the underspecification of the CEFR and to consider what this means for relating particular language examinations to the framework. It is important to recognise that the CEFR is *neutral* with respect to language and, as a common framework, it must by necessity be underspecified for all or any specific language(s). This means that specialists in the teaching or assessment of a given language (e.g. Cambridge ESOL for English) will need to determine the linguistic features which increasing proficiency in the language entails (i.e. the user/learner's competences described in Chapter 5 of the CEFR). Such features are peculiar to each language and so the CEFR must be adapted to accommodate the language in question.

ALTE's testing systems have developed alongside the CEFR over the past two decades. Many of them are now able to provide rich data and analysis to help refine the CEFR as it applies to a number of European languages. This is an important role for responsible assessment organisations to engage in and very much in keeping with the original intentions of the Council of Europe. The aim is to facilitate understanding and collaborative activities rather than to regulate or dictate to others what they should or should not do. An example of this in practice is the English Profile (EP) programme (see also *Research Notes* 33). It seeks to transpose the CEFR into English so that it becomes immediately relevant and useful to that language to curriculum designers, teachers assessment organisations and so on. Similar projects are under way in France, the Czech Republic, Germany, Spain, Georgia, Greece, Portugal and Italy to name but a few. ALTE members are involved in all cases where the country has an ALTE partner.

A major objective of the EP programme is to analyse learner language to throw more light on what learners of English can and cannot do at different CEFR levels, and to address how well they perform using the linguistic exponents of the language at their disposal (i.e. using the grammar and lexis of English). One of the main inputs to this analysis is provided by the Cambridge Learner Corpus which contains 35 million words of learners' written English from levels A1 to C2 of the CEFR. The EP research team is already providing evidence of 'criterial features' of English which are typically found in the writing of learners at the different CEFR levels. Of course this data alone does not provide an adequate sample and so part of the EP programme involves the collection of additional data from learners within the 'EP Network', including more written data and also focusing on spoken English as well (see Green forthcoming).

Examination boards and other institutions offering high-stakes tests need to demonstrate how they are seeking to meet the demands of validity in their tests and, more specifically, how they actually operationalise criterial distinctions between the tests they offer at different levels on the proficiency continuum. Cambridge ESOL, for example, has commissioned a number of 'construct volumes' in the SiLT series to assemble and present additional evidence that the examinations offered by the board are well grounded in the language ability constructs they are attempting to measure. An explicit socio-cognitive test validation framework has been developed which enables examination providers to furnish comprehensive evidence in support of any claims about the soundness of the theoretical basis of their tests (see Geranpayeh and Taylor (Eds) forthcoming in this series, Khalifa and Weir (2009), Shaw and Weir (2007), Taylor (Ed.) forthcoming, and Weir (2005)). The series develops a theoretical framework for validating tests of second language ability which then informs an attempt to articulate the Cambridge ESOL approach to assessment in the skill area under review. The perceived benefits of a clearly articulated theoretical and practical position for assessing skills in the context of Cambridge ESOL tests are essentially twofold:

- Within Cambridge ESOL – this articulated position will deepen understanding of the current theoretical basis upon which Cambridge ESOL assesses different levels of language proficiency across its range of products, and will inform current and future test development projects in the light of this analysis. It will thereby enhance the development of equivalent test versions and tasks.
- Beyond Cambridge ESOL – it will communicate in the public domain the theoretical basis for the tests and provide a more clearly understood rationale for the way in which Cambridge ESOL operationalises this in its tests. It will provide a framework for others interested in validating their own examinations and thereby offer a more principled basis for comparison of language examinations across the proficiency range than is currently available.

Cambridge ESOL is now in a position to begin a systematic and empirically based approach to specifying more precisely how the CEFR can be operationalised for English, and this in turn will lead to better and more comprehensive illustrative descriptors (particularly at the bottom and top of the scale) (Green forthcoming). In this way the CEFR will become the useful tool that it was intended to be.

In conclusion, we would like to reiterate our support for the principles and practices of the CEFR and for what we see as the main strength of the CEFR so far, its use as a communication tool (Taylor 2004). Within the common framework of levels, ALTE members have attempted to make the interpretation of examination results as transparent and meaningful as possible and

the development of functional descriptors ('Can Dos') has been useful in promoting better communication between stakeholders.

However, as noted above, it is also important to draw attention to some limitations and uses of the framework for which the CEFR was not designed. Some of these limitations were acknowledged by the original authors and some others have also been noted in the literature over the past few years (see, for example, Alderson, Figueras, Kuijper, Nold, Takala and Tardieu (2006), Fulcher (2004, 2004a), Green (forthcoming), Milanovic (2009) and Weir (2005a)).

Weir (2005a) argues that although it contains much valuable information on language proficiency and advice for practitioners, in its present form the CEFR is not sufficiently comprehensive, coherent or transparent for uncritical use in language testing. First, the descriptor scales take insufficient account of how variation in terms of contextual parameters may affect performances by raising or lowering the actual difficulty level of carrying out the target 'Can Do' statement. In addition, a test's cognitive validity – a function of the processing involved in carrying out these 'Can Do' statements – must also be addressed by any specification on which a test is based. Failure to explicate such context and cognitive validity parameters – i.e. to comprehensively define the construct to be tested – vitiates current attempts to use the CEFR as the basis for developing comparable test forms within and across languages and levels, and hampers attempts to link separate assessments, particularly through social moderation.

Weir emphasises that it is crucial that the CEFR is not seen as a prescriptive device but rather a heuristic, which can be refined and developed to better meet our needs. For the language testing constituency in particular it currently exhibits a number of serious limitations such that comparisons based entirely on the scales alone might prove to be misleading, given the insufficient attention paid in these scales to issues of validity. The CEFR as presently constituted does not enable us to say tests are comparable let alone equip us to develop comparable tests.

Taylor (2004:4) advises a cautious approach in general in using any comparative framework. She argues that:

> . . . while they promise certain benefits they can also carry inherent risks. This is because all frameworks, by definition, seek to summarise and simplify, highlighting those features which are held in common across tests in order to provide a convenient point of reference for users and situations of use. Since the driving motivation behind them is usefulness or ease of interpretation, comparative frameworks cannot easily accommodate the multidimensional complexity of a thorough comparative analysis; the framework will focus on shared elements but may have to ignore significant differentiating features. The result is that while a framework can look elegant and convincing, it may fail to communicate some key

> differences between the elements co-located within it. The result is likely to be an over simplification and may even encourage misinterpretation on the part of users about the relative merits or value of different exams.

Taylor (2004:5) concludes that:

> . . . there is no doubt that comparative frameworks can serve a useful function for a wide variety of test stakeholders: for test users – such as admissions officers, employers, teachers, learners – frameworks make it easier to understand the range of assessment options available and help users to make appropriate choices for their needs; for applied linguists and language testers frameworks can help define a research agenda and identify research hypotheses for investigation; for test providers frameworks not only help with product definition and promotion, but also with planning for future test design and development. However, we need to understand that they have their limitations too: they risk masking significant differentiating features, they tend to encourage oversimplification and misinterpretation, and there is always a danger that they are adopted as prescriptive rather than informative tools. They need to come with the appropriate health warnings!

As responsible assessment providers, ALTE members seek to provide leadership in the field of language testing, and it is important for them to address these issues explicitly. That is why organisations like Cambridge ESOL have attempted to make their stance very clear. By working collaboratively with the CEFR, the shortcomings of the illustrative scales and linguistic content can be addressed more effectively, with data being collected to enable well-informed refinements to be made as our understanding increases.

This volume therefore offers interesting insights into the application of the CEFR to language examinations and a small sample of the work that is now starting to take place. As such the papers provide a number of perspectives ranging from narrow to broad. ALTE suggested the event that led to this volume and Cambridge ESOL offered to publish the proceedings in the SiLT series precisely in order to help open up the debate on the use of the CEFR and to encourage further research along the lines illustrated here.

Michael Milanovic
Cyril J Weir
February 2010

Bibliography

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