

# Linking to the CEFR levels: Research perspectives

Neus Figueras & José Noijons (eds.)

*the Colloquium  
very well the  
sion and aims  
of the Council  
of Europe.'*



# **Linking to the CEFR levels:**

## Research perspectives

**Neus Figueras & José Noijons (eds.)**

Cito, Institute for Educational Measurement

Council of Europe

European Association for Language Testing and Assessment (EALTA)

© Cito, EALTA  
Arnhem, 2009



now you know



COUNCIL OF EUROPE    CONSEIL DE L'EUROPE

Language Policy Division  
Division des Politiques linguistiques



www.ealta.eu.org

EUROPEAN ASSOCIATION  
FOR LANGUAGE TESTING  
AND ASSESSMENT

# Preface

Cito, the Institute for Educational Measurement in the Netherlands, is proud to have sponsored the organisation of the colloquium on *Standard Setting Research and its Relevance to the CEFR* and this publication of its proceedings. Cito has much appreciated the work done by the Council of Europe in developing the CEFR. Cito has been actively engaged in the propagation of the framework and its proper use in the Netherlands and elsewhere in Europe through our participation in DIALANG and the European Survey on Language Competences, among others.

One more activity in this context has been the Cito contribution to the development of a Manual on Relating Language Examinations to the CEFR. In the Manual various methods of linking examinations to the CEFR have been described. The standardization phase in the linking process has been of particular concern to Cito. We feel that claims of links to the CEFR have not always been sufficiently substantiated. Even though the Manual gives examples of various methods of setting CEFR standards, Cito has been of the opinion that a discussion among experts could be of much use to the improvement of linking activities that many countries and institutions are now carrying out.

Cito is thankful to the European Association for Language Testing and Assessment (EALTA) to have agreed to organising the colloquium, the proceedings of which constitute this volume. Similarly Cito is thankful to the HAU, the Hellenic American Union in Athens, Greece, for having successfully hosted this colloquium.

As is borne out by the contributions in this volume, the colloquium has been a most challenging and successful event. We are certain that this publication will contribute to the dissemination of good practices in linking tests and examinations to the CEFR.

Marten Roorda, CEO Cito



# Foreword

On behalf of the Language Policy Division of the Council of Europe and EALTA we wish to express our gratitude to Cito for taking the initiative and providing generous support for the Research Colloquium held in Athens as a pre-conference event. This is the first time that the two professional bodies and the intergovernmental Organisation have joined forces in this manner. The lively interest in the colloquium and its obvious success encourage us to consider ways of continuing this co-operation.

The Colloquium served very well the mission and aims of the Council of Europe. For several decades, the Language Policy Division of the Council of Europe has worked in close co-operation with its 47 member countries and with the language education profession to develop approaches and instruments through co-operative fora and networks in order to assist in the challenging task of promoting plurilingualism in Europe. The Council of Europe has always sought and welcomed the contribution of individuals and institutions willing to share their expertise.

The Common European Framework of Reference for Languages : learning, teaching, assessment (CEFR) drew on the main developments of several decades before its publication in 2001. Its aim is to improve the quality, coherence and transparency of all aspects of language learning, teaching and assessment across Europe. In order to provide assistance in linking examinations to the CEFR, the Language Policy Division has developed and piloted a Manual for professionals in the field of assessment.

The Committee of Ministers of the Council of Europe considers it very important that linking examinations to the CEFR is carried out with due attention to the quality of the process, and stresses the need for the procedures to be well documented; results should be reported in sufficient detail and made freely available and readily accessible to all the interested parties. (Recommendation 2008 (7) of the Committee of Ministers of the Council of Europe to its member states on the use of the CEFR and the promotion of plurilingualism: [www.coe.int/t/cm](http://www.coe.int/t/cm)).

The Manual provides guidance on how to do this. Research seminars during which empirical work on linking is reported and discussed provided valuable information and concrete case studies. The Language Policy Division is very grateful to Cambridge ESOL and other ALTE members who organised events for the Council of Europe to discuss case studies on the use of the pilot version of the Manual. This event in Athens had the same goal but used a somewhat different format, and the Council of Europe is very grateful to Cito for, in addition to the organisation of the event, publishing its very interesting deliberations and findings.

EALTA's mission is broad: it aims to share professional expertise in language testing and assessment, to improve language testing and assessment systems and practice in Europe, and to engage in other activities for the improvement of language testing and assessment in Europe.

EALTA considers that, by joining forces with Cito in organising this professional event for the Council of Europe, it made a concrete contribution to the promotion of its mission. It is very gratifying that so many colleagues expressed an interest in presenting and discussing their work on relating examinations to the CEFR, a very topical and important issue for quality, coherence and transparency in language testing and examinations. We have come a long way from the seminar in Helsinki in July 2002, when the process of exploring linkage of examinations to the CEFR was launched. Together we co-operated in planning and organising the event, which launched the development of the Manual and led to the present extensive work with the Manual.

EALTA endorses the values, aims and objectives of the Council of Europe and trusts that it will soon join the group of INGOs enjoying participatory status with the Council of Europe.

The Language Policy Division of the Council of Europe welcomes EALTA's contribution and looks forward to continuing its fruitful co-operation with the Association.

Johanna Panthier  
Administrator  
Language Policy Division  
Council of Europe

Sauli Takala  
President of EALTA



# Table of contents

<b>Preface</b>	3
<b>Foreword</b>	5
<b>Introduction</b>	9
<b>Part I Approaches from theory</b>	11
Views from the expert discussants	
<b>1 Standard Setting Theory and Practice: Issues and Difficulties</b>	13
Mark D. Reckase	
<b>2 Basket Procedure: The Breadbasket or the Basket Case of Standard Setting Methods?</b>	21
F. Kaftandjieva	
<b>3 A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard-setting</b>	35
Neil Jones	
<b>4 Linking multilingual survey results to the Common European Framework of Reference</b>	45
Norman Verhelst	
<b>5 Standard Setting from a Psychometric point of view</b>	59
Gunter Maris	
<b>Part II Accounts from practice</b>	67
Reports from the practitioners	
<b>6 Relating the Trinity College London International ESOL examinations to the CEFR</b>	69
Cathy Taylor	
<b>7 Analyzing the decision-making process of standard setting participants</b>	75
Spiros Papageorgiou	
<b>8 Benchmarking of videotaped oral performances in terms of the CEFR</b>	81
Gunter Maris, José Noijons and Evelyn Reichard	

<b>9</b>	<b>Designing Proficiency Levels for English for Primary and Secondary School Students and the Impact of the CEFR</b>	87
	Karmen Pižorn	
<b>10</b>	<b>Investigating the Relationship Between the EIKEN Tests and the CEFR</b>	103
	Jamie Dunlea and Tomoki Matsudaira	
<b>11</b>	<b>Linking SQA's ESOL Higher to the CEFR</b>	111
	Rob van Krieken	
<b>12</b>	<b>Bilkent University School of English Language COPE CEFR Linking Project</b>	119
	Carole Thomas and Elif Kantarcioglu	
<b>13</b>	<b>Standard Setting for Listening, Grammar, Vocabulary and Reading Sections of the Advanced Level Certificate in English (ALCE)</b>	125
	N. Downey and C. Kollias	
<b>14</b>	<b>Jack of more Trades?</b>	131
	<b>Could standard setting serve several functions?</b>	
	Eli Moe	

# Introduction

The launch of the Preliminary Pilot version of the Manual for relating examinations to the CEFR 2003 initiated many new CEFR related projects and influenced those already in progress. The state of affairs in relation to the uses of the CEFR and associated documents in Europe, including the Manual, was analysed at the Intergovernmental Forum organized by the Council of Europe with the collaboration of Cito and the French Ministry of Education in Strasbourg in February 2007 (report accessible at [http://www.coe.int/t/dg4/linguistic/Forum07\\_webdocs\\_EN.asp#TopOfPage](http://www.coe.int/t/dg4/linguistic/Forum07_webdocs_EN.asp#TopOfPage))

At one of the preparatory meetings for the Forum, EALTA was approached by Cito, in order to jointly organise a research colloquium that would allow professionals in the field of testing to exchange views and discuss their work on relating exams to the CEFR and using the Manual. As the project fell within EALTA's mission statement *to make expertise in language testing and assessment readily available and to provide training in language testing and assessment*, and was well in line with its involvement in the dissemination of the work of Council of Europe, there was an immediate positive reaction to the proposal, which also received the encouragement of the Council of Europe.

The Research Colloquium following from this initiative was planned to maximise the opportunities to share and discuss research findings and the challenges encountered in the efforts to link language examinations to the CEFR, including standard setting, one of the most complex and controversial topics in psychometrics. A total of 12 Paper proposals were accepted, and presenters kindly agreed to send in advance their research reports for review by the panel of invited discussant experts (Neil Jones from ESOL Cambridge, Felianka Kaftandjieva from the University of Sofia, Gunter Maris and Norman Verhelst from Cito, and Mark Reckase from Michigan State University, who also gave a keynote presentation). During the Colloquium, each presentation was assigned one hour (20 minutes for presentation, 40 minutes for discussion). This format was found extremely useful and enriching, and both the Colloquium presentations and the feedback report can be accessed at [www.ealta.eu.org](http://www.ealta.eu.org).

In order to disseminate the discussions during the Colloquium the organisers agreed to publish the present book, to be organised in two parts: the first part including the texts from each of the expert discussants and the second part including short summaries of the presentations. This publication would be sponsored by Cito.

In part one, Mark Reckase's article argues the need for a theory of standard setting and presents a theoretical framework of the standard setting process, highlighting the challenges involved in the "translation of policy definitions to numerical scores". This article is followed by the contributions from the other expert discussants. Neil Jones's article reflects on the difficulties encountered in the linking process for a multilingual

project and the solution adopted, argues for the need to explore alternative approaches to linkage that may not require standard setting. Felianka Kaftandjieva's article draws the readers' attention to the importance of the choice of the standard method(s), and deals with one important criterion of the success of standard setting: the distortion of the results. She illustrates this with data from the basket procedure. Norman Verhelst has re-edited sections of the SurveyLang Inception Report for the present publication. In this article the issues in linking survey results for language tests to the CEFR are discussed. Gunter Maris' article deals with the need for relating procedures and experiences in standard setting to the psychometric literature in general, thus making it possible to use statistical methodology developed elsewhere.

The presentation summaries in part two provide an account of the important influence that the CEFR and the Manual have had in the language testing field, not only across Europe but beyond Europe (Colloquium presenters came from Scotland, Slovenia, Greece, the United Kingdom, Turkey, Japan, the USA, the Netherlands and Norway). Moreover, the summaries provide evidence of the fact that linking examinations to a standard – in this case the CEFR – is an arduous and challenging task, which requires considerable time and resources, two ingredients that have been difficult to secure for language testing in Europe so far. It is not surprising that the majority of the presentations was work in progress, a first attempt to embark on a linking process and to try out the procedures presented in the pilot version of the Manual.

It is difficult to reproduce in a book the stimulating working atmosphere, the enthusiasm, and the rich exchanges and discussions during a Colloquium, which made evident (sometimes somewhat painfully) that much more dedicated work was needed in the field of standard setting in language testing in Europe in order to be ready for international scrutiny. It is to be hoped that this book, together with the publication of the final version of the Manual at the beginning of 2009, plus the proceedings of the Conference held in Cambridge on uses of the Pilot version of the Manual in December 2007 (forthcoming) will provide both food for thought and concrete hints for further work.

A lot of people and institutions need to be thanked for having made this Colloquium possible: Cito for its funding of the expert discussants' participation and the present publication, the Council of Europe for its continuous encouragement, and the Hellenic American Union for its work prior to and during the Colloquium. Special thanks should also go to all the participants without whom this Colloquium would not have been possible. Their work before, during, and after the Colloquium was crucial in the success of the colloquium itself and in the production of the present volume. Their discipline, their intellectual generosity to volunteer ideas and to share expertise (and data, and problems, and doubts!) were the main source of the success of the Colloquium.

Neus Figueras & José Noijons, editors

Part I  
Approaches from theory:  
views from the expert discussants



# 1 Standard Setting Theory and Practice: Issues and Difficulties

Mark D. Reckase, Michigan State University

The determination of the cut score on a test that corresponds to a description of performance, a task that is often called standard setting, is one of the most important areas that fall under the general category of test theory and psychometrics. This process has this high level of importance because falling above or below the cut score on a test often has important consequences for an examinee. Unfortunately, despite the high importance of the results of standard setting, it is among the least understood areas of testing and psychometric theory. The research literature on standard setting methods and results is inconclusive, generally indicating that different standard setting methods yield different results and that the results are dependent on many of the details of implementation of the process (e.g., Jaeger, 1989, Cizek, 2001, etc.). It is my belief that the inconclusive nature of standard setting research is due to the lack of a coherent theory of standard setting that can guide the research and provide a structure for interpreting the results. The purpose of this paper is to provide the beginnings of a theory of standard setting. The theoretical framework that is provided builds on earlier work by Haertel and Lorrie (2004) and by ten years of research by ACT, Inc. supporting standard setting for the National Assessment of Educational Progress (NAEP). This work is summarized on Reckase (2000).

It is useful to begin the discussion of a theory of standard setting with definitions of a standard so that there is a common starting point. The definitions given below are from a common dictionary of the English language in the U.S. (Woolf, 1977).

*“A standard is something established by authority, custom, or general consent as a model or example: criterion.”*

*“A standard is something set up and established by authority as a rule for the measure of quantity, weight, extent, value or quality.”*

Both of these definitions indicate that standards are put in place through an active process. They are “established” or are “set up”. This means that they do not come into existence by chance. Someone calls for a standard. For the purpose of proposing a general theory of standard setting the general term “agency” will be used for those who call for the existence of a standard. The agency could be a professional organization such as those that license medical personnel, educational institutions, or a governmental unit.

When an agency calls for a standard, they usually include a statement of policy related to the standard. Such statements often indicate whether the standard is intended to be difficult to reach as in the level needed to get special recognition for academic work, or whether the standard is for the minimum requirements for all individuals. These policy statements are referred to as policy definitions of a standard in this paper. In the ideal case, these policy definitions would be very specific, directly indicating the level of performance on a test that is intended. For example, it might state that the standard is to identify the top five percent of a population. The usual case, however, is that the policy statement is somewhat vague because the agency does not know the characteristics of the test, or the test might not exist at the time that the policy definition is produced. In any case, the agency usually has some intended level for a standard and the individuals involved might be able to tell when a numerical cut score on a test does not match their intentions.

The policy definition for a standard is very important because it actually is the first setting of the standard. All other parts of a standard setting process should be consistent with the policy definition and part of the validity evidence for the results of a standard setting is the comparison of the results back to the intentions of the agency as expressed in the policy definition.

Many standard setting processes now include a step that is the elaboration of the policy definition in terms of specific statements of what persons who exceed the standard can do. In the U.S. in the educational context, these elaborations are called achievement level descriptions. In other cases, the descriptions may relate to requirements to be successful in an occupation. These elaborated descriptions should still match the intentions of the agency that calls for the standard, but they are intended to provide the specific content that is implied but not stated in the policy definition.

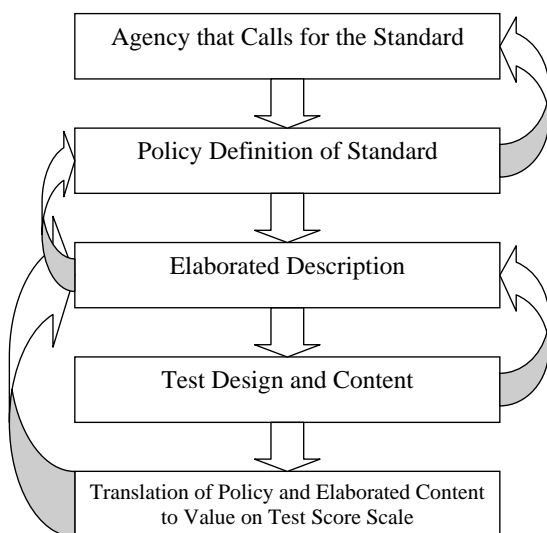
Ultimately, the implementation of the standard set by the policy stated by the agency will be operationalized as a score on a test. For this numerical cut score to be consistent with the policy definition, the test should represent the content specified in the elaborated description. If the test does not cover all of the material in the elaborated description with appropriate emphasis, then exceeding the cut score will not provide sound evidence that the examinee exceeded the standard intended by the agency. Test design and construction is very important. The best process is to design a test specifically to be consistent with a standard, but sometimes the test already exists at the time the agency calls for a standard. An important consideration then is how well the test matches the requirements set out in the elaborated description.

The final step in the standard setting process is the step that is usually called standard setting. This is the step where the policy definitions and elaborated descriptions are translated into a different language, the numerical language of the test score. Like any other translation process, it requires persons who are fluent in both the language of the policy definitions and elaborated descriptions, and the language of the test score. Because it is difficult to find people who have all of these capabilities, psychometricians and others



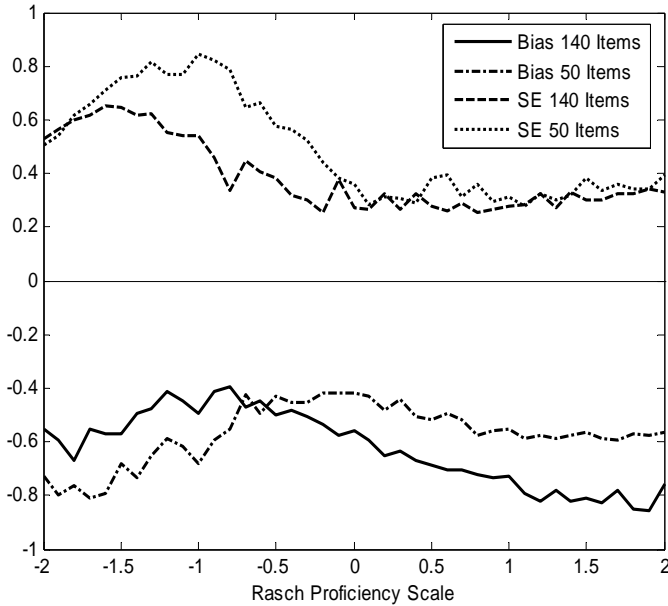
have developed processes to help with the translation process. These are the various standard setting methods such as the modified Angoff method, the contracting groups method, and others. Although it often appears that these methods are setting standards, they are really methods for facilitating the translation from verbal descriptions to numerical descriptions. The participants in the process, sometimes called panelists or judges, are really translators and they should be selected because of their skill at doing the required translations. Because this is a difficult task, multiple persons are asked to do the translations, then discuss their results and come to some consensus. Ultimately, the standard is officially set by the agency that calls for its existence and that agency needs to review all of the steps to determine if the translation accurately reflects the intentions.

The standard setting process is summarized in the diagram in Figure 1.1. This diagram shows a series of boxes with arrows moving from the agency that calls for the standards to the translation of the standards to a numerical value on the score scale for the test. There are another set of arrows that return upward to the agency that calls for the standard. These arrows indicate that the final numerical score and every step between need to be consistent with the previous step in the process. Ultimately, they need to be consistent with the intentions of the agency that calls for the standard.



*Figure 1.1 Process and Validation for Setting Standards*

This model of standard setting is controversial because it proposes that there is an intention in the standard that is specified by the policy of the agency. The hypothesis of an intended standard also means that there are standards that are inconsistent with the intentions. In other words, although the agency can not tell in advance what the numerical value on the score scale should be, they can recognize when the value that is set by a panel is different than what is intended.



*Figure 1.2 Recovery of Intended Cut Score Using the Bookmark Method*

There are many reasons why panels of judges might translate the policy definition to numerical values that are inconsistent with the intentions of the agency. One is that they do not realize that their job is not to create policy, but to translate already existing policy. Another reason is that the methodology that is used to help the translation from policy to numerical value might hinder the accurate translation rather than facilitate that translation. One example of this is given in Reckase (2006). In that research he shows that the bookmark method can underestimate intended standards if it is not carefully implemented. Figure 1.2 shows a graph of the recovered cut score compared to the intended cut score when panelists understand in detail what they are to do. The point is that if they make all of the judgments in exactly the way that they should to be consistent with policy and the characteristics of a test, they will underestimate the cut score that is intended. Figure 1.3 shows a comparison of the bookmark and modified Angoff methods that shows that the modified Angoff method does not suffer from the same problem.

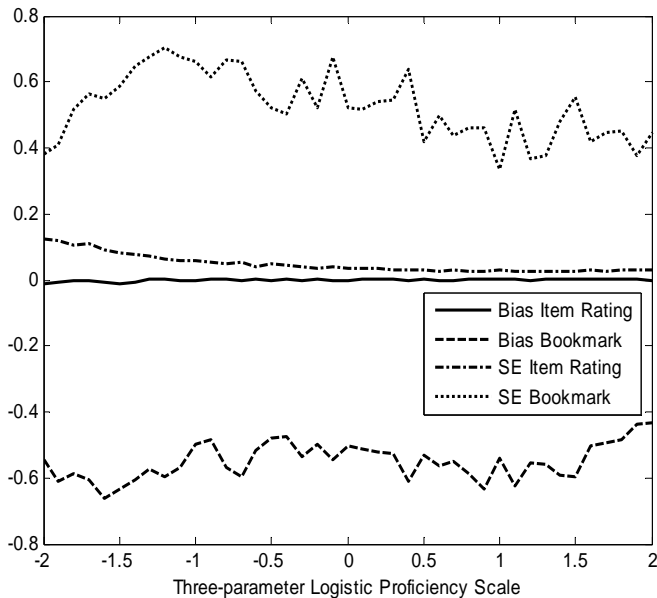


Figure 1.3 Comparison of the Bookmark and Modified Angoff Methods

Both figures give the comparison between the intended cut score and the estimated cut score for evenly spaced values on an IRT score scale. This extensive set of options is used in the research because we do not know in advance where the standard will be set so all possible values are evaluated. The results show that the bookmark method yields underestimates of the intended standard under simulated conditions and the variation in the results is high. Both of these results are not positive features of the method. Figure 1.3 shows a comparison of a simulation of the bookmark and the Angoff methods (labeled the item rating method). The figure shows that the bookmark method has greater bias and more variation than the Angoff method. Bias in this case means that the estimated cut score is different than the intended cut score. The major reason for including these results are to show that all methods are not equally good for facilitating the translation of the policy into a numerical value on the test score scale.

The model for standard setting and the evaluation of the two methods for facilitating the translation of policy to numerical cut score can be generalized to the case of the framework provided by the Council of Europe for the assessment of language proficiency. In this case, the agency is clearly the Council of Europe. That organization has called for the standards for language proficiency and has provided a policy definition for a standard. That policy definition is:

*“what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively.”*

This policy definition is not very precise, but it does give a sense of the intentions of the Council of Europe. The policy definition has been elaborated in the full description of the Common European Framework.

A complication in the case of the Council of Europe and the Common European Framework is that there are multiple tests instead of a single test that define a numerical score scale for the standard. Further, there is not one standard, but multiple standards for the dividing points for the levels A1 to C2. The fact that there are multiple tests raises the issue of how well they match the content given in the elaborated description. If they do not match well, persons who exceed the cut score will not be documented to have the skills and knowledge specified in the elaborated description. The argument for the validity of the cut score on the test is weak because of lack of content coverage.

The multiple cut scores need a more complicated translation process. There is a danger that the requirement of multiple cut scores will force a pattern of dependence in the judgments that are used in making the translations. Care must be taken to insure that the standard setting processes used to facilitate the translation of the verbal descriptions of standards to the numerical value on the score scale are not biased by the dependence between judgments.

At this point in time, the work of the Council of Europe has resulted in a policy definition for a standard and a thorough elaboration of the policy description. The challenges at this point are the multiple tests that purport to represent the elaborated descriptions and the numerous methods that are used to facilitate the translation of standards to cut scores. Current practice in the U.S. is to require that tests have evidence of alignment to the content descriptions that they purport to measure. Some of the methods used to provide evidence of alignment are of questionable value, but the basic premise is sound. Tests need to provide a representative sample of tasks that give evidence of performance on the desired content.

The issue of methods to facilitate the translation of policy definitions to numerical scores is more challenging. There has not been much research on the comparative quality of facilitating methods or on the specific situations where they might be most appropriate. The research reported here suggests that the bookmark method underestimates the standard intended by an agency and it is quite variable in the results that it provides. However, there are different ways of implementing the bookmark process. Reckase (2006) shows that some ways of using the bookmark method reduce the statistical bias in estimates of cut scores. Different implementations of the same method may not provide results that are of equal quality. It is important that the implementation of translation processes be thoroughly described and that studies be done to determine how well specific implementations support the translation of standards onto a score scale.

This paper provides only initial thoughts about a general theoretical framework for standard setting. It is not intended to be a formal evaluation of any process or test related

to the Common European Framework. It is intended to stimulate more work on developing quality procedures for facilitating the translation of standards to numerical score scales and for evaluating the alignment of tests to elaborated descriptions of policy.

# References

Cizek, G. J. (Ed.) (2001). *Setting performance standards: Concepts, Methods, and Perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Haertel, E. H. & Lorie, W. A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary Research and Perspectives*, 2(2), 61-103.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.) *Educational Measurement* (pp. 485-518). New York: Macmillan and the American Council on Education.

Reckase, M. D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice*, 25(2), 4-18.

Reckase, M. D. (2006). Rejoinder: Evaluating standard setting methods using error models proposed by Schulz. *Educational Measurement: Issues and Practice*, 25(3), 14-17.

Woolf, H. B. (Ed.) (1977). *Webster's New Collegiate Dictionary*. Springfield, MA: G. & C. Merriam Company.

## 2 Basket Procedure: The Breadbasket or the Basket Case of Standard Setting Methods?

F. Kaftandjieva, Sofia University

Standard setting aims to answer the question ‘*How good is good enough?*’ (Livingston & Zieky, 1982, p. 12). The answer to this question depends, however, to a great extent on the choice of the standard setting method, since different standard setting methods yield different cut-off scores (Jaeger, 1989; Bontempo et al, 1998).

The essential role of the choice of a specific standard setting method for the resulting cut-off score made Jaeger recommend – instead of one standard setting method in any study – to apply a combination of several standard setting methods and to establish the final cut-off score after considering all resulting cut-off scores as well as all additional information available (Jaeger, 1989, p. 500). Mainly due to the practical reasons, Cizek and Bunch hold the opposite view, asserting that it is always better to do one thing well than many things less well (Cizek & Bunch, 2007, p. 44). If this appealing advice of Cizek and Bunch is followed, the choice of the standard setting method becomes the most crucial component of the whole standard setting procedure. That is why, before attempting to answer the question: ‘*How good is good enough?*’ another question should be asked and answered and this question is: *Is the chosen standard setting method good enough?*

The aim of this article is to explore whether the Basket procedure, described in the Pilot version of the Manual (Council of Europe, 2003, p. 91), is good enough in terms of ‘*distortion of judgments*’ – an evaluation criterion proposed by Reckase (2000, 2006a). There are several reasons to focus on this issue.

**Firstly**, the Basket procedure has been applied in a number of European projects aiming to link language tests to the CEFR, but despite its frequent implementation there is no published research to analyze its quality and to compare the method with other standard setting methods. This lack of supporting research is a serious drawback (Reckase, 2006b) and it contravenes one of the four evaluation criteria, suggested by Norcini & Shea (1997, p. 43), for the standard setting methods applied in a credentialing setting.

**Secondly**, according to Reckase and Bay, the Angoff yes/no method and its modifications result in “*statistically biased estimates of the standards*” (Reckase & Bay, 1998, p. 12). The Basket procedure might also be expected to lead to biased estimation of the standards due to its similarity with the Angoff yes/no method. Applying the method without checking for possible bias would cast some serious doubts on the validity of the established cut-off scores.

**Thirdly**, there are a number of evaluation criteria available (Hambleton & Powell, 1983; Berk, 1986; van der Linden, 1994; Norcini & Shea, 1997; Kane, 1998; Reckase, 2000; Cizek et al, 2004; Schafer, 2005; Reckase, 2006a; Hambleton & Pitoniak, 2006; Cizek & Bunch, 2007), but most of them do not concern the quality of the standard setting method itself but its implementation in a specific testing situation (Berk, 1995; Kane, 1998).

By contrast with the mainstream of existing evaluation criteria, the ‘distortion of judgments’ criterion is based on the idea that a good standard setting method is the method which, properly implemented, allows judges to recover their intentions (Reckase, 2006b, p. 15). This criterion assumes error-free perfect judgments by a judge and closely corresponds to Standard 4.21. (AERA et al., 1999, p. 60). The argument that the real life is far from perfect (Schulz, 2006, p. 5) is not a valid argument against this criterion, because what does not work well in perfect conditions is unlikely to work well in the imperfect world, either.

In other words, the first step is the choice of the method and only after that comes the proper implementation of the chosen method. Following this logic, without ignoring the importance of other evaluation criteria, this study will focus only on the criterion for minimal distortion of judgments and will answer the question whether the Basket procedure meets this criterion.

## **Instrument**

The test used in this study is an English Reading comprehension test, targeted at B1/B2 CEFR levels. It was developed by S. Takala, N. Figueras and F. Kaftandjieva especially for the needs of the EALTA pre-conference workshop ‘Standard Setting in Practice: How to cut the mustard?’ (15-17 June 2007, Barcelona).

The test consists of 41 items with selected response (multiple-choice, true/false and matching items) and in terms of difficulty level, item format and length the test resembles to a great extent language tests used to assess language proficiency in the field of reproductive skills (listening, reading, grammar and vocabulary).

The test was administered to 334 examinees from Finland and Catalunya in 2007 and Table 1 presents basic test and item statistics. The results in Table 1 show that the test has acceptable quality in terms of reliability. The examinees from the first sub-sample perform slightly better than the examinees from the second sub-sample, but the difference between the mean scores in both cases (raw score and theta score) is less than one standard error of measurement.



Table 2.1 Test Statistics for the test of Reading Comprehension

Statistics		Sub-sample 1 (Finland)	Sub-sample 2 (Catalunya)	Total
Sample		212	122	334
Items		41	41	41
Item Difficulty	Min	33%	17%	29%
	Mean	66%	63%	65%
	Max	98%	98%	97%
Item Discrimination	Min	0.21	0.09	0.20
	Mean	0.39	0.39	0.38
	Max	0.59	0.80	0.63
Raw score	Min	8.00	11.00	8.00
	Mean	27.16	25.62	26.60
	Max	41.00	38.00	41.00
	SD	7.07	6.91	7.04
	$\alpha$	0.86	0.87	0.86
	SEM	2.65	2.49	2.63
Theta score	Min	-1.03	-0.80	-1.03
	Mean	+0.69	+0.46	+0.69
	Max	+3.87	+1.74	+3.87
	SD	0.69	0.61	0.67
	$\alpha$	0.86	0.88	0.87
	SEM	0.26	0.21	0.24

One of the advantages of the Basket procedure is that it allows standard setting to be done without the application of IRT. It is an important feature of the method, which allows its broader application. On the other hand, although, in principle, it is possible to check the requirement for minimal distortion of judgments in terms of the raw score (number correct), in this case the IRT approach was preferred.

That is why, in addition to the statistics concerning the raw scores, Table 1 presents the corresponding statistics also for the IRT based  $\theta$ -scale. The IRT model used in this particular case is the one parameter logistic model – OPLM (Verhelst & Glas, 1995). This model was chosen, because in contrast with the Rasch model, the OPLM fits the data quite well for all test items ( $p_i > 0.03$  for  $i = 1, 2, 3, \dots, 41$ ) as well as for the test as a whole (Global fit:  $R_{1c} = 161.36$ ;  $df = 160$ ;  $p = 0.46$ ).

The OPLM is an extension of the Rasch model, which allows items to differ in their discrimination and the discrimination indices are imputed as known integer constants in contrast with the two-parameter logistic model where they are estimated (Verhelst et al., 1995, p. 1-2). As a result, the OPLM often fits data which the Rasch model would not fit due

to its strong assumptions. In this specific case the main reason for the Rasch model misfit was exactly the wide range of item discrimination indices (between 0.20 and 0.63 in terms of item-test correlation).

The frequency distributions of examinees and items on the  $\theta$ -scale are shown in figure 2.1.

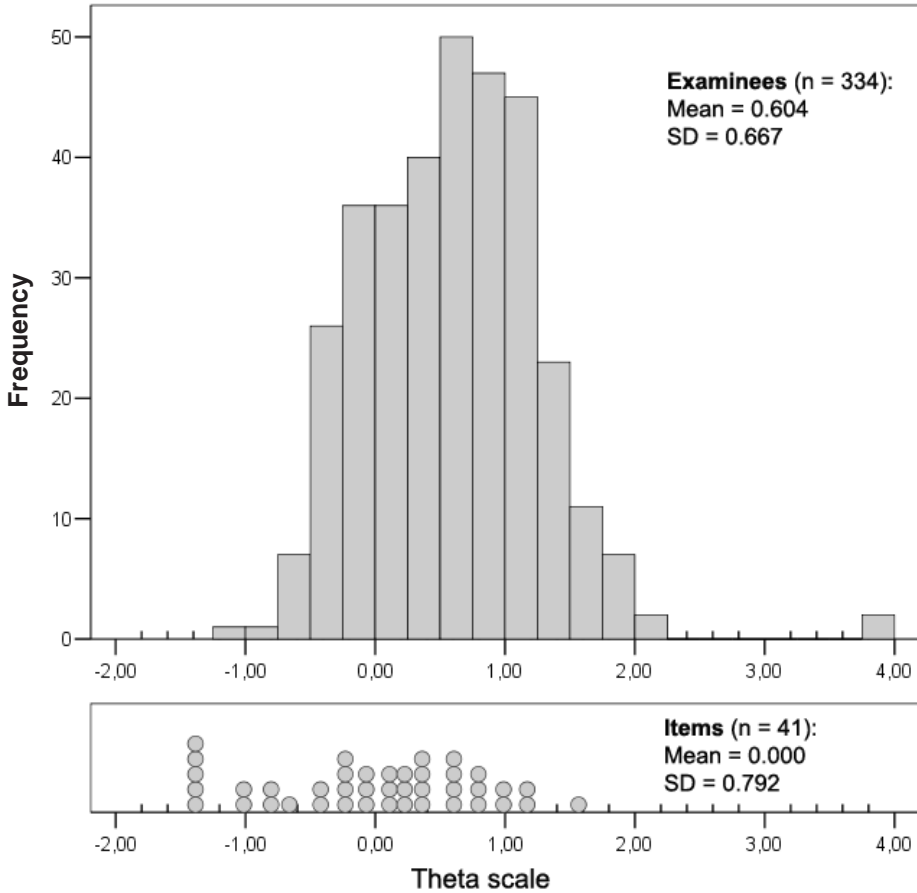


Figure 2.1 Frequency Distributions of Examinees & Items ( $\theta$ -scale)

The graphs in figure 2.1 show that as a whole the test is relatively easy for this sample of examinees with good coverage of items in the middle range (-1; +1) of the  $\theta$ -scale.

### Basket procedure

The Basket procedure (Council of Europe, 2003, p. 91) is a standard setting method developed especially for the needs of the linkage between language tests and the CEFR (Council of Europe, 2001). In its initial version (Kaftandjieva et al, 1999; Verhelst & Kaftandjieva, 1999), applied in the first phase of DIALANG, the **judgment task** requires

the judge to decide for every item *whether a test-taker on level X should be able to answer it correctly* and these decisions have to be made for each proficiency level.

This judgment task requires a yes/no type of answer and in this respect is similar to the Angoff yes/no method (Cizek & Bunch, 2007, pp. 88-92). The major difference between the two judgment tasks is in the conceptualization of the target examinee: an examinee on level X in the Basket procedure and a minimally proficient examinee in the case of Angoff yes/no method. In this respect the judgment task of the Basket procedure (in its initial version) is much more similar to the judgment task of the Jaeger's iterative two-choice method (Jaeger, 1989, pp. 494-495).

To clarify the difference between the judgment tasks for the Basket procedure and the Angoff yes/no method, a useful parallel can be made with the two main examinee-centered standard setting methods: the Contrasting groups method and the Borderline group method (Cizek & Bunch, 2007, pp. 105-117). The Basket procedure is analogous to the Contrasting groups method, because both require yes/no decisions to be made about the membership of an item/examinee to a certain proficiency level/group. The Angoff yes/no method on the other hand is analogous to the Borderline group method, because in both cases a judge has to identify items/examinees that belong to the borderline group (items which could be answered correctly by a borderline person or examinees who can be qualify as borderline test takers).

The initial format of the judgment task in the Basket procedure leads to multiple rounds of standard setting in case there is a need to set multiple cut-off scores (one per level), which is time consuming. To overcome this limitation, the judgment task was modified and in its current format the judgment task requires for each item in the test the judges to answer: *“At what CEF level can a test taker already answer the following item correctly?”* (Council of Europe, 2003, p. 91). The reference to the Framework (CEFR) in this formulation does not limit the application of the method in the context of different frameworks as long as the levels of proficiency form an ordinal scale.

To set the **cut-off score** between two consecutive proficiency levels ( $X_{i-1}$  &  $X_i$ ), only the items, judged to be answered correctly on the levels preceding level  $X_i$  ( $\leq X_{i-1}$ ), are taken into account. In case of raw score reporting (Classical Test Theory approach) the cut-off score between levels  $X_{i-1}$  &  $X_i$  is equal to the number of items judged to be answered correctly at levels below level  $X_i$  ( $\leq X_{i-1}$ ). If test results are reported in  $\theta$ -score (OPLM approach) the cut-off score between levels  $X_{i-1}$  &  $X_i$  in  $\theta$ -scale is equal to the value of  $\theta$  corresponding to the weighted raw score of items judged to be answered correctly at levels below level  $X_i$  ( $\leq X_{i-1}$ ). The weights in this case are equal to item discrimination indices used in OPLM calibration (in this case – integers varying between 1 and 3 with a geometric mean equal to 1.52 and a sum of 67). In other words, if someone answers correctly all 41 test item, his/her weighted raw score will be equal to 67 if this specific IRT model is used.

There is a need for some clarification about the interpretation of the cut-off score itself: the question is whether the cut-off score  $C_{X_{i-1}/X}$  is the upper bound of level  $X_{i-1}$  or the lower bound of level  $X_i$ . Since, in terms of the raw score, the cut-off score is equal to the number of items judged to be answered correctly by a test-taker who is at level  $X_{i-1}$  or below, it seems more sensible to consider the cut-off score  $C_{X_{i-1}/X}$  as the upper bound of  $X_{i-1}$  level. The reason for this is that at this point of the raw score a person should be able to answer correctly the items assigned to levels below level  $X_i$ , but still will not be able to answer correctly items assigned to upper levels ( $>X_{i-1}$ ) according to the judgment task. This interpretation of the cut-off score as an upper bound of  $X_{i-1}$  level differs from the interpretation given in the Manual for relating language examinations in the CEFR (Council of Europe, 2003, p. 91) where the cut-off score  $C_{X_{i-1}/X}$  is considered (without justification) to be the minimum requirement for level  $X_i$  – in other words, its lower bound. The argument, presented here, is in concordance with the logic of the judgment task and cut-off score establishment and justifies the decision for the cut-off score between two consecutive proficiency levels ( $C_{X_{i-1}/X}$ ) to be treated as the upper bound of level  $X_{i-1}$ .

The main advantage of the Basket procedure is its **practicability**. According to Berk, “*practicability refers to the ease with which a standard-setting method can be implemented, computed, and interpreted*” (1986, pp. 143-144). The fact that the Basket procedure has been applied in a number of European projects aiming to link language tests to the CEFR supports the statement that this standard setting method meets all four practicability criteria suggested by Berk. Berk himself gives another support by analogy by rating the Angoff yes/no method (which is similar to the Basket procedure) as meeting all four criteria.

Another advantage of the Basket procedure is that its application **does not require IRT** modeling, which extends the area of its possible implementations and in this sense is closely related to the practicability of the method.

The judgment **task complexity** is another evaluation criterion (Reckase, 2000, p. 51-52), which the Basket procedure probably meets. The item mastery method (Kaftandjieva et al, 1999; Verhelst & Kaftandjieva, 1999), which is a standard setting method based on the same judgment task, but uses different methodology for cut-off score establishment, was rated by Reckase as a method with low task complexity (Reckase, 2000, p. 54). There are also indications that judges usually perceive the task as easy, but there is a need for research supporting the validity of this statement.

One of the main **disadvantages** of the Basket procedure is that it is **based purely on judgments** without taking into account the empirical information. Most of the early test-centered methods have the same drawback, which made Berk (1986) classify them as **Judgmental** in contrast with the examinee-centered methods which, according to his terminology, are **Empirical-Judgmental** methods. Nowadays the predominant view is that “*normative information needs to be made part of the process for judges to anchor their absolute judgments with an understanding of current levels of performance of students and likely consequences*” (Linn, 2003, pp. 14-15).

To overcome this shortcoming, many of the modern test-centered standard setting methods have multiple rounds in which judges are provided with additional information – feedback, item statistics, impact data and possibility to discuss their judgments. Another approach is to incorporate the empirical data as an integral part of the standard setting process as it is done in the case of the Bookmark method or the Item Descriptor Matching Method (Cizek & Bunch, 2007, pp. 155-207) when the items are presented to the judges in ordered (in terms of difficulty) item booklets.

Multiple rounds and ordered item booklets are widespread procedures in standard setting, but they also have their own weaknesses. A different methodology is used in two other versions of the Basket procedure: the Item mastery method (Kaftandjieva et al, 1999; Verhelst & Kaftandjieva, 1999), which is in fact the first version of the Basket procedure, and the Cumulative compound method (Kaftandjieva & Takala, 2004). In these two standard setting methods, the integration of judgments and empirical data is done not during the judgment session, but at the stage when the cut-off score(s) is(are) established. In this sense, the parallel between these two methods (Item mastery method and Cumulative compound method) and the Contrasting groups method is much closer than the parallel between the Basket procedure and the Contrasting groups method. In fact, the Cumulative compound method can best be described as a test-centered Contrasting groups method.

In contrast with all these possible approaches for integrating judgments with empirical data, the Basket procedure, as it is described in the Pilot version of the Manual (Council of Europe, 2003, p. 91), seems to be the only contemporary standard setting method that is based purely on human judgments only.

Another **disadvantage** of the Basket procedure, which is shared by all methods in this family, is that there are circumstances when **the cut-off score cannot be determined**. For example, if a judge assigns all items at levels equal or above  $X_i$ , it will be impossible to set a cut-off score between levels  $X_{i-1}$  &  $X_i$ . It might not look as a very likely situation, but experience shows (Moe, 2008) that in practice it happens regularly, especially with narrowly focused tests and in cases when the intended cut-off score is close to the ends of the score range. For instance, in the standard setting workshop (Barcelona, 2007) the Basket procedure was applied with 34 judges to set-up three cut-off scores (A2/B1; B1/B2 and B2/C1) for the same Reading comprehension test, presented here. The ratings of 7 out of 34 judges, however, did not allow setting either the A2/B1 cut-off score or the B2/C1 cut-off score. The Bookmark method has the same problem (Reckase, 2006), but in both cases (Basket procedure and Bookmark method) the proper solution has not been found yet.

### **Minimal Distortion of Judgments**

Let us imagine that a judge knows in advance the cut-off score he/she wants to set up applying a particular standard setting method. Using Reckase's terminology this known-in-advance cut-off score will be called – **intended cut-off score (ICS)**. The question is whether the judge will be able to recover this intended cut-off score if he/she:

- follows strictly the standard setting procedure,
- has all information available and
- makes perfectly consistent judgments with the ICS.

If the result of the application of the standard setting method in these perfect conditions differs from the intended cut-off score, “...then its credibility for estimating the ICS would be questionable when errors in judgments are added to the process” (Reckase, 2006, p. 5).

The hypothesis is that the Basket procedure will not meet the criterion for the minimal distortion of judgments due to its similarity with the Angoff yes/no method, which is potentially biased (Reckase, 2000, p. 51; Cizek & Bunch, 2007, p. 94; Hambleton & Pitoniak, 2006, pp. 440-441).

To illustrate the evaluation of the Basket procedure in terms of the criterion for the minimal distortion of judgments, let us consider  $\theta_{ICS} = -0.09$  as the intended cut-off score. The reason to choose this specific  $\theta$  value as the ICS is that it is the cut-off score between levels B1 and B2, set on the basis of the ratings of the 34 judges participating in the standard setting workshop (Barcelona, 2007) applying the Basket procedure to link the same test to the CEFR.

The attempt to recover ICS = -0.09 implementing the Basket procedure means to assign to level B1 (or below) all items which can be answered correctly by a person whose ability level is less or equal to -0.09. The rest of the items should be assigned to levels equal or above level B2.

In probabilistic terms, the statement that a test taker at level X can answer this item correctly means that the probability of correct answer at level X is greater than 0.5. There are altogether 17 items with probability of correct answer  $> 0.5$  at  $\theta = -0.09$ . The weighted raw score of these 17 items is equal to 23 and corresponds to  $\theta_c = -0.37$ . This value ( $\theta_c = -0.37$ ) is lower than the intended cut-off score ( $\theta_{ICS} = -0.09$ ) and the difference between the two scores ( $\theta_c - \theta_{ICS} = -0.28$ ) in absolute terms is greater than the standard error of measurement at  $\theta = -0.09$  ( $SEM_{(\theta = -0.09)} = 0.21$ ). This result is in support of the hypothesis that the implementation of the Basket procedure in perfect conditions might lead to biased estimation of the cut-off scores.

To check for potential bias at the other points of the  $\theta$  scale, the same analysis was done for all possible ICS in the interval between -2.94 and +2.82. There was, however, a problem with the estimation of the cut-off scores at the ends of this interval due to the fact that for  $\theta_{ICS} < -1.39$  the probability of correct answer for all items was  $< 0.5$  and for  $\theta_{ICS} > +1.5$  the probability of correct answer was  $> 0.5$  for all items. Facing the same problem with the Bookmark method, Reckase set the cut-off scores arbitrarily below (above) the easiest (most difficult) item in the test (Reckase, 2006, pp. 9-10). Instead of arbitrariness, ‘the avoidance approach’ was applied in the current study – meaning that the analysis was done only for the range of  $\theta$  scale in which the cut-off scores can be estimated (-1.39; +1.5). The avoidance approach does not solve the problem and in this sense it is not better than

the arbitrary approach. If, however, a similar analysis precedes the real standard setting, its results might be very useful in deciding what is the most appropriate standard setting method for that specific situation.

The results of this analysis for the Basket procedure applied in perfect conditions to set the standards for that particular test are presented in figure 2.2.

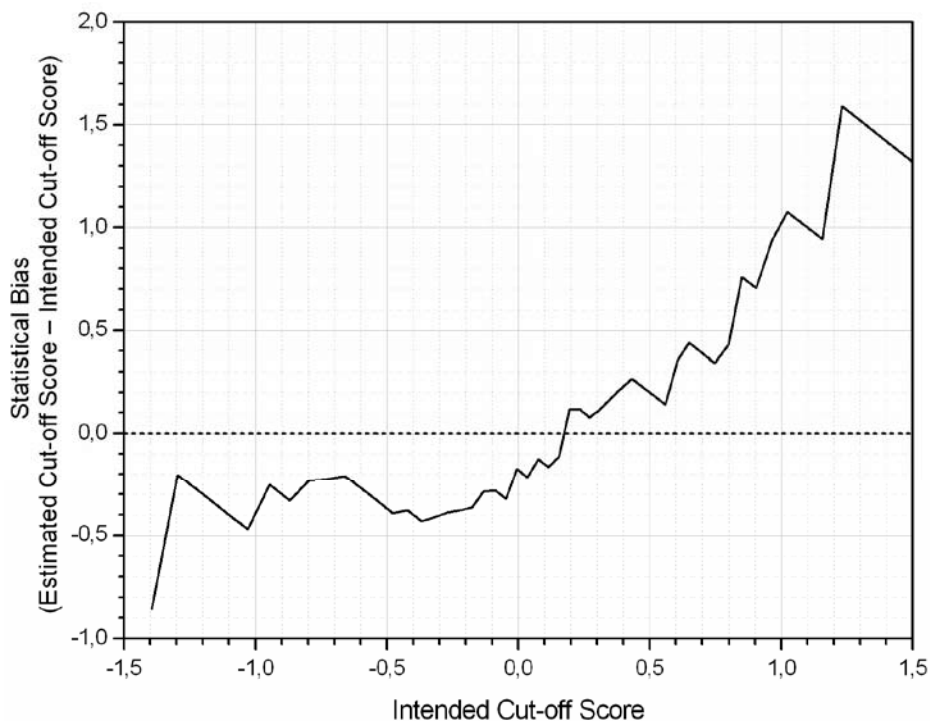


Figure 2.2 Basket Procedure: Discrepancy between Estimated Cut-off Score & Intended Cut-off score

The solid line on figure 2.2 shows the difference between the estimated cut-off score (based on the Basket procedure) and the intended cut-off score for each possible intended cut-off score (X-axis) in the interval of  $\theta$  where the estimation can be done. If the Basket procedure were able to recover the intended cut-off scores, the solid line would coincide with the horizontal dashed line at  $y = 0$ . The graph, however, shows a large discrepancy between the two cut-off scores (intended and estimated) with a clear tendency for **underestimation at the left end** of scale range and **overestimation at the right end**. The average absolute difference between the two scores (intended and estimated) is 0.45 and this difference is bigger than the standard error of measurement for the corresponding intended cut-off scores in 66% of the cases. Even more, for 12 out of the 52 intended cut-

off scores (23%), for which the calculations were done, the difference was larger than two standard errors of measurement with the extreme of 5.6 SEM for  $\theta_{ICS} = +1.2$ . In fact, the discrepancy between the two scores is within acceptable limits (within one standard error of measurement) only in the very narrow interval of  $\theta$  between 0.00 and +0.56. Outside this interval, there is a dramatic difference between the estimated cut-off score and the intended cut-off score.

These results confirm the hypothesis that the Basket procedure, implemented in perfect conditions, cannot recover the intended cut-off score and produces biased estimations (underestimation at the left end of the scale and overestimation at the right end). These findings cast serious doubt on the credibility and defensibility of standards set on the basis of the Basket procedure and suggest the application of the method to be limited only to the cases of low-stake tests or as a complementary standard setting method only.

## Conclusion

In a situation which Kane felicitously describes as “*an embarrassment of riches*” (Kane, 1998, p. 137), when there is a great number of standard setting methods to choose from, and knowing that they usually lead to different cut-off scores (Jaeger, 1989; Bontempo et al, 1998), the question whether a certain standard setting method is good enough becomes crucial.

The aim of this study was to explore whether the Basket procedure is good enough in terms of the minimal distortion of judgments and the answer to this question is “No”.

Someone might argue that this conclusion is based on a single test only and hence generalizations are questionable. This is true in principle, but since the results of this study make the use of the Basket procedure questionable then whenever the Basket procedure is applied in a real testing situation, evidence should be provided in order to refute this conclusion and to justify the choice of the method.

Another interesting question is whether the existing modifications of the Basket procedure have the same problem and which of them is better in terms of the minimal distortion of judgments. The answer to this question goes beyond the scope of this study, but other research shows that the distortion of judgments for the Item mastery method and the Cumulative compound method is smaller than for the Basket procedure and usually within one standard error of measurement (Stoyanova, 2008).

It will also be interesting to explore what will be the distortion of judgments for the Basket procedure when different IRT models (if they fit) are applied as well as when the Classical test theory approach is used. In addition to the pure research interest, such exploration is important and should become an integral part of the internal validation of standard setting, based on any standard setting method applied under specific conditions in a concrete testing situation.



In the urge to link (preferably quickly) the existing language tests in Europe to the CEFR, the quality of this link has been often overlooked. It is about time, however, to focus on the quality of the standard setting methods applied as well as the quality of their implementation, because the established cut-off scores matter and they matter for over five million examinees who annually sit and take high-stake language tests in Europe and pay for it.

# References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Berk, R. (1986). A Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests. *Review of Educational Research*, 56(1), 137-172.
- Berk, R. (1995). Something Old, Something New, Something Borrowed, a Lot to Do! *Applied Measurement in Education*, 1995, 8(1), 99-109.
- Bontempo, B., Marks, C. & Karabatos, G. (1998). *A Meta-Analytic Assessment of Empirical Differences in Standard Setting Procedures*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 1998.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA : Sage.
- Cizek, G., Bunch, M. & Koons, H. (2004). An NCME Instructional Module on Setting Performance Standards: Contemporary Methods. *Educational Measurement: Issues and Practice*, 23(4), 31-50.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Council of Europe. (2003). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF)*. Manual – Preliminary Pilot Version. Strasbourg: Language Policy Division.
- Hambleton, R. & Pitoniak, M. (2006). Setting Performance Standards. In: *Educational Measurement*. (4th ed.) R. Brennan (Ed.), American Council on Education / Praeger, Westport, CT, 433-470.
- Hambleton, R. & Powell, S. (1983). A Framework for Viewing the Process of Standard Setting. *Evaluation & the Health Professions*, 1983, 6(1), 3-24.
- Jaeger, R. (1989). Certification of Student Competence. In: *Educational Measurement* (3rd ed.), Ed. by R. Linn, Washington DC: American Council on Education and Macmillan, 485-514.

Kaftandjieva, F. & Takala, S. (2002). *Relating the Finnish Matriculation Examination English Test Results to the CEF Scales*. Paper presented at Helsinki Seminar on Linking Language Examinations to Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Retrieved July 18, 2008 from <http://kielet.hkkk.fi/kievi/Kaftandjieva.htm>

Kaftandjieva, F., Verhelst, N. & Takala, S. (1999). *DIALANG: A Manual for Standard setting procedure*. (Unpublished).

Kane, M. (1998). Choosing Between Examinee-Centered and Test-Centered Standard-Setting Methods. *Educational Measurement*, 1998, 5(3), 129-145.

Linn, R. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives*, 11(31). Retrieved September 16, 2003 from <http://epaa.asu.edu/epaa/v11n31/>

Livingston, S. & Zieky, M. (1982). *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Princeton, NJ: ETS.

Moe, E. (2008). *How to make judges' life easier?* Paper presented at the EALTA Pre-Conference Research Colloquium "Linking tests to the CEFR", Athens, May 2008.

Norcini, J. & Shea, J. (1997). The Credibility and Comparability of Standards. *Applied Measurement in Education*, 10(1), 39-59.

Reckase, M. (2000). A Survey and Evaluation of Recently Developed Procedures for Setting Standards on Educational Tests. In: Bourque, M. and Byrd, S. (Eds.) (2000). *Student Performance Standards on the National Assessment of Educational Progress: Affirmations and Improvements*. Washington, DC: National Assessment Governing Board. 43-70.

Reckase, M. (2006a). A Conceptual Framework for a Psychometric Theory for Standard Setting with Examples of Its Use for Evaluating the Functioning of Two Standard Setting Methods. *Educational Measurement: Issues and Practice*, 2006, 25(2), 4-18.

Reckase, M. (2006b). Rejoinder: Evaluating Standard Setting Methods Using Error Models Proposed by Schulz. *Educational Measurement: Issues and Practice*, 2006, 25 (3), 14-17.

Reckase, M. & Bay, L. (1998). *Analysis of Methods for Collecting Test-based Judgments*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego; CA; April 14-16; 1998).

Schafer, W. (2005). Criteria for Standard Setting from the Sponsor's Perspective. *Applied Measurement in Education*, 18(1), 61-81.

Schulz, E. (2006). Commentary: A Response to Reckase's Conceptual Framework and Examples for Evaluating Standard Setting Methods. *Educational Measurement: Issues and Practice*, 2006, 25 (3), 4-13.

Stoyanova, F. (2008). *Criterion-Referenced Educational Testing: Standard Setting*. Sofia: Daniela Ubenova. (In Bulgarian).

van der Linden, W. (1994). *A Conceptual Analysis of Standard Setting in Large-Scale Assessments*. Research Report 94-3. Twente University, Enschede (Netherlands): Faculty of Educational Science and Technology.

Verhelst, N. & Glas, C. (1995). The One Parameter Logistic Model. In: G. Fischer and I. Molenaar (Eds.), *Rasch Models: Foundations, recent developments, and applications*, New York: Springer Verlag, 215-238.

Verhelst, N., and Kaftandjieva, F. (1999). *A Rational Method to Determine Cutoff Scores (Research Report 99-07)*. Enschede, The Netherlands: University of Twente, Faculty of Educational Science and Technology, Department of Educational Measurement and Data Analysis.

Verhelst, N., Glas, C. & Verstralen, H. (1995). *One Parameter Logistic Model (OPLM)*. Arnhem: Cito.

### 3 A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard-setting

Neil Jones, Cambridge ESOL

#### **Standard setting and the CEFR: the problem**

I was surprised, though pleased, to be invited to participate in this colloquium as a discussant, because for some years now I have been expressing an essentially sceptical view of standard setting for the purpose of constructing a multilingual proficiency framework – that is, as advocated in relation to the CEFR. My position dates from 2003, when I started work on the construction of a similar framework for a new set of UK qualifications, called *Asset Languages*. This project was offered as a case study of applying the pilot versions of the manual for aligning exams to the CEFR (Council of Europe 2003). When it came to presenting the case study I had to confess that it was more about where we had *not* followed the pilot version of the manual than about where we had followed it. I also emphasized that this was not in itself a judgment on the manual, but rather reflected the development procedures that we had been forced to adopt by the multilingual scope (twenty-five languages) and tight schedule of that particular project.

What the *Asset Languages* project did impress on me was the need to look for ways of taking a high-level, top-down view of the process of constructing such a framework, with respect to its vertical dimension of progression through levels, and its horizontal dimension of alignment and comparability across languages. This requires us to find methods of working explicitly with these two dimensions, rather than dealing with each language and level separately in the optimistic belief that decisions made at micro-levels will lead to the emergence of a coherent whole.

Clearly, there are many different aspects to implementing a top-down model in a test development project and pursuing the (finally unachievable) goal of perfect comparability across languages. The manual is a valuable resource, which undoubtedly facilitates the adoption of a top-down model, with the CEFR providing the coherent framework to which each language, and each group of language learners, may be aligned. (Incidentally, I prefer to speak of aligning language learners to the CEFR rather than language tests, because in the end the process concerns the validity of tests, and validity concerns the inferences we make about the learners who take them. If a test for young learners is difficult to align to the CEFR, it is because the argument that links young learners' language performance to the CEFR is currently harder to construct.) However, I was not fully convinced by the treatment of standard-setting in the pilot version of the manual, and I will now try to explain why.

Firstly, I find the use of terminology slightly idiosyncratic. The term ‘standard setting’ is used in Chapter 5 to refer to task-centred approaches and to objective tests, while learner-centred approaches are treated as ‘external validation’ in Chapter 6. This suggestion of logical and temporal priority – that task-centred standard setting is an essential first step, and that learner-centred approaches are an option for some later validation stage – does not seem to reflect the treatment of these two broad approaches in the literature, where both are simply referred to as standard setting.

Secondly and chiefly, I feel that the use of task-centred standard setting approaches in constructing a multilingual framework is a misapplication of techniques to a situation where their underlying premises and justification do not hold.

I can make this clearer by describing what I would call a *classical* standard setting context – the one in which many of these approaches developed – and contrasting it with our purpose in relation to the CEFR. In this way I can identify the issues which I believe require our particular attention.

Let’s take as the classical context the professional licensure exam: say, for example, a one-hundred item MCQ test for nurses. We can characterise this context in terms of the following premises:

1. The judges and candidates are members, or prospective members, of a specific professional community;
2. The test tasks relate to discrete items of professional knowledge;
3. The judges are qualified to say which items a practitioner should master;
4. Hence the notion of ‘minimal competence’ has substantive meaning;
5. The buck stops with the judges, who are responsible to the public. Judgments are not ‘correct’, only defensible in terms of balancing the interests of the candidate nurses and the public whom they will serve.
6. The frame of reference is the profession and its stakeholders, and no judgments have implications outside this frame of reference;
7. The judges’ professional and cultural background (for their practice is culturally embedded) impacts on their decisions and actually reinforces their validity (within that culture).

At the colloquium Mark Reckase made the point that judges should not be seen as *setting* standards, but rather as converting into a cutoff score an intended standard previously specified by some higher agency. Relating this to the ‘classical’ situation described above, I don’t think it substantively changes my contention that the standard is defined within the frame of reference defined by the profession.

Anyway, the CEFR context clearly differs in several important respects. Listening and Reading are skills: tests do not simply measure discrete atoms of knowledge, but attempt to tap hidden mental processes (violating premises 2 and 3 above). Hence we are dealing with an indirectly observable continuum of ability: the notion of minimal competence, or any discrete level of competence, is hard to pin down (violating premise 4).

The frame of reference is languages across Europe, and so *all* judgments have implications which extend beyond the immediate context of a particular test or language (violating premise 6).

Judgments *can* and *must* aspire to be 'correct' in the sense of *consistent* with other judgments being made within the wider frame of reference (violating premise 5). Therefore the culturally-determined nature of judgments, far from reinforcing their validity, becomes a serious threat to it (premise 7).

This last point in particular presents the major challenge for aligning standards across languages. Clearly, the whole purpose of the CEFR is to provide that practical point of reference that enables a common understanding of levels. But level descriptors are not wholly concrete or definitive. They require interpretation, and our default expectation must be that different countries' interpretations will be culturally determined (in a broad sense) and therefore may differ. This is, of course, not just a hypothetical problem, but is a recognized current practical issue which is now beginning to be addressed, most notably in the important multilingual benchmarking event held by CIEP in Paris in June 2008.

In this section I have argued that the assumptions or premises which justify orthodox task-centred standard setting approaches are violated in the case of linking language tests to the CEFR. It is necessary to look at the problem in a different way.

### **Absolute and comparative judgment: rating and ranking**

If the CEFR's frame of reference takes in all European languages then clearly the correctness of a standard set for any language can only be evaluated by comparison with other languages. Instead of attempting absolute judgments about the level represented by a score on a Reading or Listening test, or a sample of performance in Writing or Speaking, we need to think in terms of comparative judgments: is this Reading task in language X harder or easier than this task in language Y? Is this sample of Speaking in language X better or worse than this sample in language Y? The basic act of judgment in a multilingual frame of reference is thus not *rating*, but *ranking*. This reflects a general principle that constructing a framework is logically a two-stage process: first we construct a common measurement scale, and second we set standards.

I can try and make this point more clearly by offering an analogy with measuring and interpreting temperature. Historically the first step was to construct a practical measuring instrument – a thermometer. The next step was to calibrate it – that is, put a numbered scale on it. It evidently made sense to devise a standard scale and ensure that all thermometers reported using it. Today celsius has become the standard scale for most purposes. Only at this point did it become practical to develop *interpretations* of points on the scale. We have been able to develop and share a sophisticated understanding of how seriously to treat varying degrees of fever precisely *because* our measurement instruments are aligned to the same scale.

To relate this back to our multilingual framework: it makes logical sense *first* to align tests across languages to the same scale, and only then to develop interpretations – i.e. set standards. Those interpretations will then apply equally to all the aligned languages. Of course, what makes logical sense is not always possible in practice - it certainly wasn't in the case of *Asses Languages*, and neither is it in the case of the CEFR, where so much has already taken place. However, what I propose here could contribute to the current iterative process of progressive approximation to the intended meaning of the CEFR levels.

By focussing on comparative judgments – ranking – we can achieve the alignment of language tests and performances to the same scale. We should find this an easier, more practical task because human beings are much better at making comparative judgments than absolute judgments. Bramley (2005) quotes Laming (2004), who goes so far as to say: “There is no absolute judgment. All judgments are comparisons of one thing with another.” It also addresses the more fundamental question. In my understanding, the question “Is my B1 your B1?” is first a question about equivalence, and only second a question about the meaning of B1.

And if we can answer this question we are already much better placed to answer the second question – the one about interpretation, or standards. Obviously, a comparative approach cannot remove the need for standard setting at some stage, but by placing it at a logically later stage – after the alignment of languages to a common scale – it dramatically reduces the scope of standard setting. The standard is set once but applies equally to all aligned languages. Subsequent languages can be aligned to the same framework by a relatively simple comparative exercise. There is no need – in fact it is not possible – to do standard setting separately for each such language, because the act of alignment applies the standard already set.

Thus we can conclude that the logic of a multilingual framework is such as to severely constrain the freedom of judgments relating to individual languages. If we accept this then there follow further possible conclusions for the methodology of framework construction. Concerning objectively-marked tests of Reading and Listening, it remains a problem for standard setting to establish meaningful cutoffs on what are essentially continuous measurement scales relating to indirectly observed mental processes. For these skills in particular it is comparability of measures which is paramount. If we can develop a measurement scale, or appeal to some existing one, which defines levels rationally in terms of the way they relate to substantive learning gains, likely learning hours between levels, or the definition of accessible learning targets, then we can argue that this scale could be applied by default across languages. As North (2006) suggests, the CEFR has developed out of a concept of levels which are appropriate for broad groups of learners at particular stages in their learning career, and taken together define a progression which makes sense as a ‘learning ladder’. Taylor and Jones (2006) describe the development of the Cambridge ESOL levels and their relationship to the CEFR in similar terms.



This was the approach adopted with the Reading and Listening scales for *Asset Languages*, where we adopted as a prototype or template the common scale upon which the Cambridge ESOL levels have been calibrated. That is, experience of working with these scales, which of course depends on an item banking, IRT scaling methodology, gave us a useful expectation of how a scale for similarly tested skills should look. I've written about this idea elsewhere in relation to scaling the CEFR levels for objective tests (Jones 2005); it is mentioned here just to reinforce the point that in a multilingual framework freedom of standard setting judgment is very severely constrained, one of the constraints being the proportional placement of levels on a measurement scale developed using a particular kind of response data.

### **Data collection and analysis for a ranking approach**

I have stated that ranking allows us to align languages to a common scale, which in turn allows us to set the same standards for all the aligned languages. I will now look at methods that we can use.

Bramley (2005) reviews comparative approaches. The earliest of these is Thurstone's paired comparison method (Thurstone 1927), which is based on the idea that the further apart two objects are on a latent trait, the greater the probability of one of them 'winning' a comparison. Thus from a set of dichotomous judgments (e.g. of 'better' or 'worse') one can estimate not simply an ordinal ranking, but the relative location of each object on an interval latent trait scale. Thurstone's model can be implemented in different ways, of which the most computationally tractable is a Rasch formulation (Andrich 1978).

However, a practical problem found by Bramley and others using paired comparisons is the repetition and sheer number of paired judgments required. A ranking approach, where more than two objects are compared, is thus an attractive alternative. One analysis approach is to decompose ranking data into paired comparisons, although because these are of necessity self-consistent they lack independence and thus violate Rasch model assumptions, exaggerating the length of the measurement scale. Alternatively rankings can be used as categories in a Rasch partial credit model. Here the top-ranking object 'scores' 1, the second 2 and so on, for each judge involved. Bramley shows that the methods produce highly correlated results. Linacre (2006) reviews different methods of analysing rank-ordered data.

Bramley (2005) treats the case of a National Curriculum test of Reading attainment for pupils aged 14, and of equating performance on test versions from one year to the next. He distinguishes standard setting from what he calls *standard maintaining*: a comparative approach is used here to attempt to apply the standard from a previous year to the current year. The objects of comparison were scripts containing pupils' responses to short-answer Reading questions. This was thus a comparison of pupils' performance. It allowed the pupils from the two years to be aligned on a single ability scale, from which equivalent cutoff scores for the two test versions were estimated by linear regression of marks on ability. Results from this ranking study were found to agree well with an equating based on different information.

The multilingual benchmarking conference organised by CIEP at Sèvres in June 2008, subsequent to the Athens standard-setting colloquium, also focussed on performance, this time the skill of Speaking. Two kinds of data were collected. At the conference itself judges rated video performances against the CEFR, using a similar methodology to earlier such events conducted for French, German and Italian between 2004 and 2006, but with the difference that ratings were elicited in a “cascade” design using English and French as “anchor” languages: working in one group (on English and French), then in two and then three parallel subgroups, each dealing with 3 languages (i.e. English, French and one other). Prior to the conference ranking data were collected from the same judges, using a specially-developed web-based platform which allowed them to view samples and record their ranking by dragging samples to re-order them in a list. The allocation of samples for the ranking exercise was such as to ensure that each judge rated in two languages, and that there was linkage in the data across all samples and languages.

At the time of writing only a preliminary analysis has been completed, which shows good correlation between the two approaches, and which should further our understanding of the relationship between them. A positive outcome from analysing data of this kind would confirm the utility of the ranking approach and also provide the basis for aligning subsequent languages to the CEFR in the manner I proposed above, by alignment to a core of well-equated and standard-set languages via a relatively simple comparative exercise, with no further role for standard setting.

There remains the issue of whether a comparative approach can be made to work for the task-centred case of Reading or Listening as measured by objectively-marked items. This is the most difficult case – the one where, I have argued, the standard setting approach proposed in the pilot version of the manual is least convincing. The focus here is not on samples of performance, but on comparison of test items across languages. Apart from a small workshop I conducted at the ALTE conference in Vilnius in November 2007 I have no data to base a claim on. However, I would be reasonably hopeful that a comparative approach can be made to work.

Ranking items by difficulty is on the face of it a simpler task than, for example, stating the probability that a hypothetical borderline candidate might get an item correct, which appears to involve some highly abstract reasoning and reference to an entity (the hypothetical candidate) which may be very differently interpreted by judges. It is true that judges have been found to be not very good at estimating the relative difficulty of test items, hence the practice in some standard-setting methods of providing the item difficulties to them. But a procedure could be explored in which judges would be observed ranking items by difficulty for a single language as well as across two languages. The cross-language comparison could be done in various ways, with knowledge of the relative difficulties of none, or one, or both of the item sets. Outcomes could be correlated with item calibrations from empirical data to derive indices of probable accuracy with respect to the single-language and by extension to the cross-language case. Thus we should be able to find ways of estimating standard errors of an alignment, interpretable in ways that the

standard errors provided by standard-setting are not. And finally, the direct comparison of items remains a simpler, more concrete cognitive process than those demanded by orthodox task-centred standard setting methods.

## Conclusion

I have proposed that a ranking approach offers a practical way to align different languages and tests to a common scale, and that it is logical to do this as a separate step prior to standard setting. This priority reflects our fundamental concerns in building a multilingual proficiency framework: first, to establish equivalence of levels, secondly to assign meanings to them.

This priority may not be obvious to the majority of groups concerned with aligning a test or a set of exam levels to the CEFR, because they are interested in a single language. However, as we begin to establish a core of credibly-linked languages, with the associated benchmark exemplars of performance, and calibrated sets of tasks for objective skills, both the feasibility and the compelling arguments in favour of adopting a comparative approach will become clearer.

There is further work to do developing and validating comparative methodologies for aligning performance skills and objective test items, and deriving indices enabling us to evaluate outcomes. There are undoubtedly wider and more fundamental issues concerning the nature of comparability, which are highlighted by the approach but already lie at the very heart of a frame of reference like the CEFR. On what basis *can* younger learners be compared with adults, or formal learning compared with informal acquisition through immersion? These are not questions to address here, but from observing benchmarking events it seems that certain conventional ground rules have to be agreed. Perhaps the nature of these rules still requires better articulation and theoretical justification. My position here is simply that if we accept the utility and practicality of aligning different learning contexts to the CEFR then the comparative approach is a good way of doing so.

A major benefit of the comparative approach is that it is, at least on the face of it, robust against variation in how judges from different backgrounds understand the intention of the CEFR levels. Severity or lenience can have no impact on a ranking – unless it is applied differentially across languages. While there is no a priori reason for believing this to be likely in the case of Europe's most widely used languages, we cannot dismiss the possibility entirely. A desire to preserve the national language against corruption by neighbouring closely-related languages, perhaps linked to a political wish to limit immigration, is one possible scenario where standards might be set too high. Perhaps you can think of others. But this would be a problem for any methodology, and a multilingual approach with multinational participation should be as robust against such effects as any method can be.

A comparative approach to the CEFR necessarily constrains the scope of standard setting, making it in many cases neither necessary nor possible. Partly this is because priority is given to scale construction. It occurred to me during the colloquium that some of the cases

presented, even though their focus might be on a single language and even a single level, were indeed using standard setting methods to address what were properly scaling issues. As the pilot version of the manual itself makes clear, having a valid and reliable approach to test construction is a pre-requisite for addressing standards, because without it the standard may simply fluctuate from session to session. Thus effort put into developing an item-banking, IRT-based approach to test construction may pay more dividends than the same amount of effort devoted to repeated standard setting. That is, the standard-setting case should be distinguished from the standard-maintaining case.

Also, I believe the final version of the manual will make clearer the importance of that wider range of validation activities focussing on the competence of learners, and based on information from a range of sources beyond the test itself, that constitute 'standard setting' in its wider meaning. This is the area covered under 'external validation' in the current pilot manual. Can-do studies relating to target domains, for example, may be illuminating (and can also be treated in a comparative way). It would be good if the focus of standard setting activity around the CEFR were to be widened out in this way.

# References

- Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, 2, 449-460
- Bramley, T. (2005). A Rank-Ordering Method for Equating Tests by Expert Judgment. *Journal of Applied Measurement* 6(2), pp202-223
- Council of Europe (2003). *Relating language examinations to the CEFR. Manual; Preliminary Pilot Version*. Retrieved from: [http://www.coe.int/T/E/Cultural Co-operation/education/Languages/Language Policy/Manual/default.asp](http://www.coe.int/T/E/Cultural%20Co-operation/education/Languages/Language%20Policy/Manual/default.asp)
- Jones, N. (2005). Raising the Languages Ladder: constructing a new framework for accrediting foreign language skills. *Research Notes No 19*. Cambridge ESOL. Retrieved from: [http://www.cambridgeesol.org/rs\\_notes/offprints/pdfs/RN19p15-19.pdf](http://www.cambridgeesol.org/rs_notes/offprints/pdfs/RN19p15-19.pdf)
- Laming, D. (2004). *Human judgment: The eye of the beholder*. London: Thomson
- Linacre, J. M. (2006). Rasch Analysis of Rank-Ordered Data. *Journal of Applied Measurement*, 7(11), 129-139
- North, B. (2006). *The Common European Framework of Reference: Development, Theoretical and Practical Issues*, paper presented at the symposium 'A New Direction in Foreign Language Education: The Potential of the Common European Framework of Reference for Languages', Osaka University of Foreign Studies, Japan, March 2006.
- Taylor, L. and N. Jones (2006). Cambridge ESOL exams and the Common European Framework of Reference (CEFR) . *Research Notes* 24. Retrieved from: [http://www.cambridgeesol.org/rs\\_notes/offprints/pdfs/RN24p2-5.pdf](http://www.cambridgeesol.org/rs_notes/offprints/pdfs/RN24p2-5.pdf)
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 3, 273-286



## 4 Linking multilingual survey results to the Common European Framework of Reference

Norman Verhelst, Cito & SurveyLang

### 4.1 Introduction

This chapter is an adaptation of relevant sections in the Inception Report of the SurveyLang consortium pertaining to the problem of linking survey results for language tests to the Common European Framework of Reference (CEFR).

The European Council in Barcelona in 2002 called for further action to ‘improve the mastery of basic skills, in particular by teaching at least 2 foreign languages from a very early age’, and for the ‘establishment of the linguistic competence indicator’. This decision arose from the current lack of data on the actual language skills of pupils and the need for reliable data to measure the progress towards this new objective. A detailed approach for the creation of a European survey was outlined in the Communication from the Commission to the Council “Framework for the European Survey on Language Competences”.

The survey should cover tests in the first and second most taught official European languages in the European Union;

- from a representative sample of pupils in education and training at the end of ISCED level 2 (or from ISCED 3 if a second foreign language is not taught before);
- test performance should be interpreted with reference to the scales of the Common European Framework of Reference for languages (CEFR);
- the indicator should assess competence in the 3 language skills which may be assessed most readily (i.e. listening comprehension, reading comprehension and writing).
- Instruments for testing in these 3 competences should be developed, taking into account the previous experience and knowledge in the field at international, Union and national level.

The SurveyLang Consortium successfully tendered to conduct this large-scale language survey. SurveyLang is an international team made up of experts in the fields of test development, sampling and data collection, as well as in educational measurement, cognitive psychology, research design, scaling, and data analyses. SurveyLang represents a range of countries. The SurveyLang consortium consists of (in alphabetical order):

- Centre international d’études pédagogiques (CIEP)
- Gallup
- Goethe-Institut
- Instituto Cervantes

- National Institute for Educational Measurement (Cito)
- University of Cambridge ESOL Examinations (Cambridge ESOL)
- Universidad de Salamanca
- Università per Stranieri di Perugia

In section 4.2 the general principle of the linking will be discussed and it will be explained why we need to have recourse to methods involving Item Response Theory. The main point of that section, however, is to make clear that a linking procedure is not just a mechanical routine, which when applied carefully, leads to valid results.

Section 4.3 discusses two problems related to the construct validity of the constructed scales. One treats the dimensionality problem, the other the cross language equating of the constructed scales.

In section 4.4 the proposed methods of linking will be presented. The methods are chosen from the ones presented in the manual for linking examinations to the CEFR (Council of Europe, 2009).

In section 4.5 the problem of validation is discussed, i.e., the problem of finding independent corroborating evidence which shows that the linking process is trustworthy.

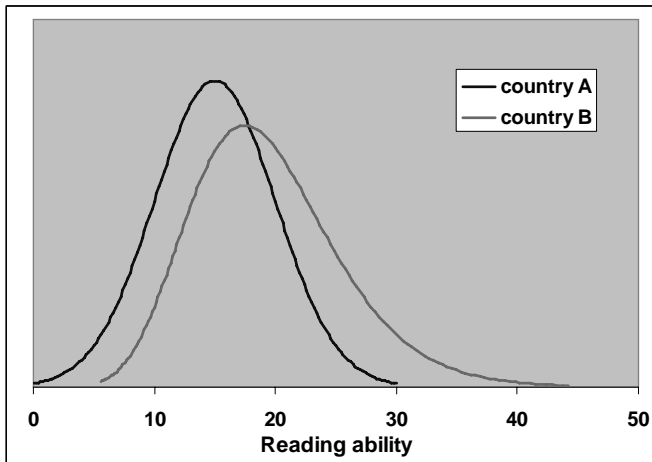
## **4.2 Measurement and Standard Setting**

To cover a fairly broad domain of content and a quite large range of proficiency, much more test material has to be developed than can be administered in a reasonable time and meaningfully to each student. This implies that the use of a single standardised test to be administered to all sampled students is not possible for the projected survey. Or, equivalently, each student will answer to only a rather restricted subset of the test material. The technical term for such a set up is testing in an incomplete design.

An immediate consequence of using incomplete designs is that raw scores obtained on different tests are not immediately comparable, even not in the case where each student takes an equal number of items or tasks, because some subsets may contain more difficult material than others. To arrive at meaningful comparisons the different subsets used in the survey must be equated in some way.

A powerful method of equating is the use of Item Response Theory (IRT). The basic assumption in this approach is the existence of an underlying continuous scale (which we might conceive of as an ability or proficiency scale), while the answers to items and tasks are conceived as indicators. The precise way in which this indicating function is to be understood is usually quite involved, but the essential characteristics of this approach are (1) that the higher the ability, the more probable it is that a correct response will be given; (2) items and tasks also have an 'ability value' on the same scale as the test takers. In the most simple model (the Rasch model), this value is usually called the difficulty of the item.





*Figure 4.1 Basic outcome of the survey*

By using this approach, items and students can be represented on the same scale, essentially independent of the particular subset of items that have been administered to different groups of students.

The primary outcome from the survey is graphically displayed in Figure 4.1. For each country the basic outcome is an estimate of the ability distribution. We comment on the essential characteristics of this Figure:

- The numbers along the horizontal axis are expressed on an interval scale; the unit and origin of this scale is arbitrary.
- The figure refers to the reading ability for a single language. For each language, one can construct such a figure, but for different languages unit and origin are arbitrary, such that the scales for two languages cannot be directly compared.
- From the figure one can derive directly interesting comparisons between the two displayed distributions for the countries A and B:
  - The average ability in country B is higher than in country A;
  - The variance in country B is larger than in country A;
  - The distribution in country A is symmetric; in country B it is skew to the right.

Note that this is the finest grained scale possible on which one can report since it is continuous. From a measurement point of view, linking to the CEFR amounts to a coarser way of categorising students than using a continuous scale. This is exemplified in Figure 4.2, for the same distributions as in Figure 4.1. The three vertical lines cut the horizontal axis into four pieces. The left vertical line defines the boundary between alleged CEFR levels A1/A2. It cuts the horizontal axis at the value 12, implying that all students having an ability value not larger than 12 are categorized as A1.

The procedure to arrive at placing the vertical lines is called standard setting, and the cut-off values themselves are called the (performance) standards. Once the standards are set, it

is a routine procedure to determine for each country the proportion of students belonging to each of the CEFR categories (A1 or lower), A2, B1 and (B2 or higher). For example: the proportion of students in country A being at level A1 or lower corresponds to the area under the leftmost curve from zero to the leftmost vertical line.

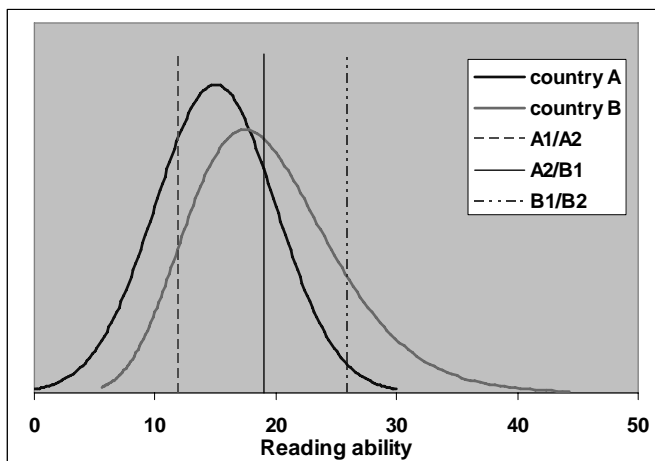


Figure 4.2 Basic outcome and standard setting

The foregoing procedure is simple enough in principle, but it is based on a number of assumptions, and the validity and usefulness of its results will depend on the truth of these assumptions.

Here is a short list of possible problems. The remainder of this chapter is a description of how these problems will be tackled in the survey.

Is there such a thing as reading ability for a particular language, which can be represented as a single continuous line?

If the first question is answered affirmatively, can we conceive of reading ability in English (as a foreign language) as essentially comparable to reading in French or German? Note that this does not imply that students should have the same reading ability in English as in French. Concretely, the question is whether we can represent the same person by two points on a single scale, such that the relative position of these points with respect to each other represent differences in reading ability of the same person in the two different languages. More generally, can we make test performances of the same student in different languages comparable to each other.

The third question is more technical in nature: by which procedure do we arrive at the placement of the vertical lines in Figure 4.2? And, very important: is the procedure such that the results (the cut-off points on the horizontal axis) can be considered as plausible values to distinguish between the CEFR levels. This will be discussed in the next section. Notwithstanding all efforts to arrive at plausible standards, good scientific practice

requires external validation, i.e., independent evidence that corroborates the results of the standard setting: do we arrive at essentially the same conclusions if we use independent sources of information.

### **4.3 Questions of Construct Validity**

Does there exist something like a unidimensional scale (for some skill in some language)? Although the mathematical elegance of many IRT models might invite people to believe they represent the truth, the models themselves and everything that can be deduced from them by correct mathematical reasoning only have the status of a hypothesis, and the basic ideas underlying these models must therefore be subjected to quite severe testing.

Well elaborated IRT models and the accompanying software will provide also tools to test the validity of the model assumptions. In the context of the survey two aspects are of utmost importance. The first has to do with the construct within target languages; the second has to do with comparability across languages.

#### **4.3.1 Undimensionality of the measured constructs**

It may be the case that the items for a single skill, e.g. reading, are so different in content, domain and other respects that the observed responses cannot be described formally by invoking a single (unidimensional) ability, but that more dimensions have to be invoked. A standard technique to determine this is factor analysis; exploratory as well as confirmatory. Special attention will be given to the distinction between a more communicative dimension versus a more formal one (grammar, lexis). This exploration will be carried out for the five target languages separately.

An important feature of a survey is to find a reasonable compromise between scientific rigor and communicative reporting for the main target audience, the national governments. Therefore it may be necessary to ignore to some extent fine grained subtleties in the main report, while in other respects it may appear to be important to accentuate the finer structure of the data. It seems reasonable, for example, that a report can be made where the distribution of the reading skill is described for different countries, and that this description captures the main components of a possible multidimensional construct, without necessitating endless nuances at the risk of making the report barely accessible or useful.

This does not imply, however, that deviations from unidimensionality can simply be ignored. There are at least two aspects where attention is due to a possible multidimensional structure of the data. One is the problem of differential item functioning (DIF) which is discussed in more detail elsewhere<sup>1</sup>. The other is important with respect to standard setting, and will be addressed in the next section.

---

<sup>1</sup> The discussion on DIF is not included in the present publication

### **4.3.2 Cross language test equating**

Having found evidence that the constructs within the five target languages are reasonably well defined, for example, that it makes sense to speak about reading ability or proficiency in the five target languages and that a scale has been defined in these five languages, there remains the problem of the relations between these five reading scales. In the construction of the test material, care has been given to make the framework common to the five target languages, and the test material itself will be constructed in such a way that the materials for the five languages parallel each other as much as possible, such that it can be expected that the constructs across languages are very similar if not equivalent. Although such content based arguments may increase confidence in the cross language comparability of the tests, they are not sufficient to warrant numerical equivalence of test scores or related concepts (estimated ability).

Defining the scales within the five languages is based on psychometric analyses, which cannot be generalised in a simple way to a common scale for all languages together. Such a generalisation would require (a) that students are tested in at least two languages (which is not permitted by the terms of reference) and (b) that these students are equally proficient in the two languages they are tested in, an assumption that is not testable nor realistic.

Therefore student responses cannot be used to establish a common scale of reading ability, say, for the five target languages. Comparability, however, can be reached using judgements of experts who are able to judge the relative difficulty of tasks or the relative merits of performances in two or more languages. To arrive at a common scale for writing for two languages, a number of judges will be asked to rank order a set of students' works, about half of them in one language, and the other half in the other language. By a suitable design and suitable IRT modelling the mixed set of works can be scaled on a common scale. (See Bramley, 2005). As will be seen in Section 4.5 this equating will contribute to a large extent to the empirical validation of the standard setting procedures.

## **4.4 Standard setting procedures**

The procedures for standard setting will be different for the receptive skills (Reading and Listening) and for the productive skill of writing. For all procedures, the big steps described in the manual for relating language examinations to the CEFR (Council of Europe, 2009) will be followed: specification, familiarisation, standardisation and standard setting proper will be applied in compliance with the manual.

### **4.4.1 Test specification**

The test specification follows automatically from the testing framework and from the parallel development of test material in the five test construction teams for the five languages. Although it may not be realistic to discuss each and every item or task with representatives from all participating countries, the framework itself will be subject to international consultation and discussion, so as to lead to a high degree of acceptance of the tests.

#### **4.4.2 Familiarisation**

As the standard setting is one of the key issues of the survey it is important that all participating countries are involved in this process. In principle, for each of the five target languages a panel will be composed consisting of at least one representative of each country testing that language.

The familiarisation will comprise two aspects: familiarisation with the test and the test specification and familiarisation with the CEFR itself. The time needed for familiarisation will of course depend on the familiarity of the participating panel members with the CEFR.

#### **4.4.3 Standardisation of judgements**

This is especially important for productive skills. As the writing tasks have to be marked by judgemental procedures, the markers will be instructed by the use of clear exemplars. In the standardisation of the judgments for standard setting the same exemplars will be used, as local benchmarks. At the same time it will be investigated if internationally approved benchmarks are available for supporting the judgments during standard setting.

#### **4.4.4 Standard setting procedure for Writing**

For writing, a student centered method will be used. The kernel of the procedure consists of allocating students to one of the CEFR levels on the basis of all the writing tasks they have completed. This method is more attractive than the classic student centered methods (as the contrasting groups method or the borderline group method) because all panel members make a judgement on the same set of completed tasks, while the latter methods usually are based on a single judgment per student, and therefore these judgments are not open to discussion. Details on the statistical analysis of the given judgments can be found in Cizek & Bunch (2007) and in chapter 6 of the manual (Council of Europe, 2008).

In piloting the method, it will be investigated whether the use of benchmarks and exemplars during the standard setting method proper is helpful for increasing the consistency of the responses.

#### **4.4.5 Standard setting procedure for Reading and Listening**

For the two receptive skills, a test centered method of standard setting will be used, viz. the Cito-variation of the bookmark method. The section in the manual treating this method is reproduced here (text in the box).

The bookmark method<sup>2</sup> may get more complicated if the items do not discriminate equally well (which is more often the case than not). A simple example with two items is displayed in Figure 6.3., where the dashed curve represents the best discriminating item. The two curves represent item response functions: they relate the latent ability (horizontal axis) to the probability of obtaining a correct response (vertical axis).

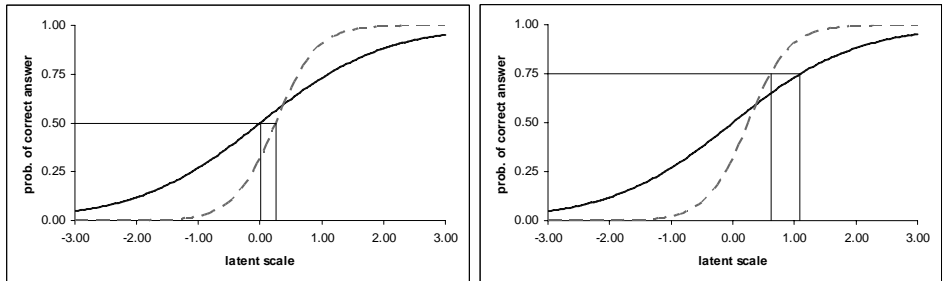


Figure 6.3 Items with unequal discrimination

If one uses the bookmark method with  $RP^3 = 0.5$  (left-hand panel), the dashed item will have a higher page number (being presented as the most difficult of the two) in the OIB than the other one, while with an RP of 0.75 (right-hand panel), the reverse holds: the dashed item will now appear as the easiest of the two. This illustrates the fact that ‘difficulty of an item’ is not a simple concept, and presenting the ordering of the difficulties by a simple number may confuse panel members.

The method developed at Cito aims at presenting in a graphical way difficulty and discrimination values of all items in a single display. Consider the least discriminating item in Figure 6.3: for  $RP = 0.5$  the required ability is 0.; for  $RP = 0.75$ , the required ability is about 1.1. One could designate a chance of 50% to get an item correct as ‘borderline mastery’, while a chance of 75% correct could be called ‘full mastery’. To go from borderline to full mastery the ability must increase from 0. to 1.1, and one can display this graphically in a figure like Figure 6.4, an item map (for 16 items) that contains information about difficulty and discrimination of each item. Each item is represented as a piece of line, stretched horizontally. The left end corresponds to the difficulty parameter of the item ( $RP = 0.5$ ), and the length is indicative for the discrimination value: the longer the line, the less the item discriminates. The right end corresponds to a higher RP, 0.75 or 0.80, say. The display is constructed in such a way that the left ends of the item lines increase as one goes from bottom to top. One should take care that the lines are properly identified, such that panelists can associate each line clearly with an item in the test.

<sup>2</sup> In the bookmark method the items are presented to the panel members in order of increasing difficulty in a so-called ordered item booklet (OIB). Panel members set their subjective standards by inserting a bookmark at the page where the items (at a predetermined level of mastery) seem to jump to a higher level, e.g., the first

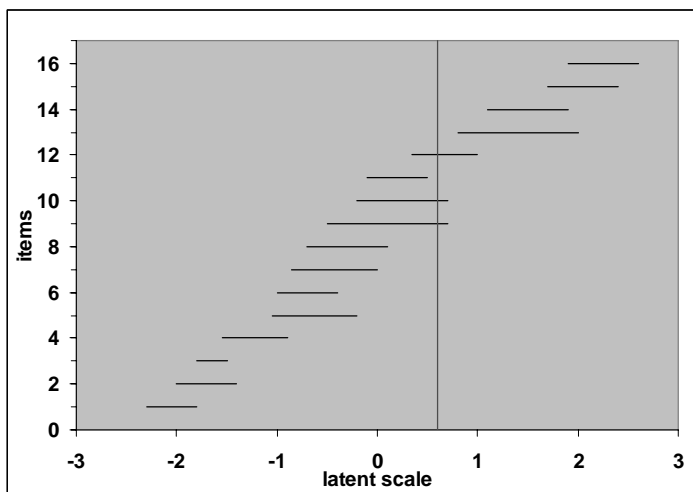


Figure 6.4 Item map, indicating difficulty and discrimination

The vertical line symbolizes the provisional standard of a panel member, and by drawing this line (or holding a ruler) the panel member can quickly have an overview of the consequences of his/her decision. In the example the proposed standard implies full mastery of the items 1 to 8 and of item 11; for the items 9 and 10 there is almost full mastery; for item 12 borderline mastery has been reached, and for the items 13 tot 16 borderline mastery is not reached at all.

To apply the method, the panel members can be asked to draw a vertical line, or to give a numerical value that corresponds to the location where the vertical line touches the horizontal axis in the figure (which is 0.6 in the example)

Notice that from Figure 6.4 it cannot be deduced in any way how the distribution of latent abilities in the target population looks like. To avoid all associations with, e.g. a normal distribution, it may be advisable to change the scale values which are displayed along the horizontal axis in the Figure to a convenient scale having no negative values, and an easy to understand unit. For example, adding 8 to all numbers displayed along the axis in the figure, and the multiplying by 10 will make the numbers range from 50 to 110, avoiding interpretations in terms of percentages, and being fine grained enough to require provisional standards expressed as whole numbers. After the standard setting is completed, the obtained standards can easily be transformed back to the original scale, and standards in the score domain are determined in the same way as in the bookmark method (see Section 6.8.) [End of quotation from the manual]

bookmark may indicate the transition from level A1 to level A2 of the CEFR.

<sup>3</sup> RP stands for response probability, and indicates the mastery level alluded to in the previous note.

#### **4.4.6 Multiple rounds**

In modern standard setting methods, it is customary to ask judgments from the panel members in several rounds. Typically, the first round follows the training phase of the procedure. After the first judgment, the results are summarized and panel members are invited to discuss possible discrepancies in judgments. This kind of information is called normative. After discussion, a second judgment is elicited, summarized, and followed by a new round of discussions where usually information is given with respect to the score distribution in the target population, to make panel members aware of the consequences of their judgments.

As it is the intention to work with very heterogeneous panels, and as information about the distribution of the abilities in one or more countries is politically very sensitive, this kind of impact information is probably not very suitable in the survey. The content of the second discussion round will be discussed in the next section.

#### **4.5 Empirical validation**

Although the manual treats the different phases in linking examinations to the CEFR as linearly ordered in time (specification – standardisation of judgments – standard setting – empirical validation), sticking to such a linear order too strictly may be disappointing in the end when, for example, it appears in the empirical validation that the panel members in the standard setting have been too lenient or too strict, compared to an independent criterion.

Instead of arranging the whole process as linear in time, in the survey it will be attempted to run the processes as much as possible simultaneously, and to allow multiple independent sources of information to influence each other.

##### **4.5.1 Sources of information**

Essentially, five different sources of information relevant to the linking to the CEFR will be made available:

In the development of the test material items and tasks will be related closely to the CEFR. Specifically this means that for the receptive skills item specifications will refer to descriptors in the CEFR, will relate to exams offered by SurveyLang partners at different CEFR levels, and can be anchored to these at the trialling and pretesting stages. This will ensure validity and general appropriacy of level. For the productive skill (Writing), the scoring rubrics will be formulated in terms of the CEFR and rater training will centre on these. Practically, this implies that from the test specification itself, a provisional standard setting follows (see the manual, chapter 6).

Answers from the sampled students are important input for standard setting procedures. This is clearly the case for student centred methods (the body of work method to be used for Writing), but they are also essential for the so-called test centred method to be used for the receptive skills. In the older student centred methods (like the Angoff method), the validity of the procedure depends critically on the ability of the panel members to have a



clear view on the relative difficulty of the items, to have a constant (within panel members as well as between panel members) conception of a 'borderline person' and to give realistic probability statements. In the method we propose to follow (see the preceding section) the task for the panel members is conceptually much easier, although items not only differ in difficulty but also in discrimination. The ordering of the items in difficulty is given (in a single graphic overview), differences in discrimination are easily associated with features of the graph, and the difficult concept of a borderline person does not appear in the procedure. In contrast, panel members have to set a provisional standard (and thus define a borderline person in pure psychometric terms), and can see the consequences (in probabilistic terms) for each item. To provide such information, one needs of course responses from students. Figure 6.4 (see section 4, in the box) is in a sense a concise summary of the test behavior of the sample of students. Therefore, the term test centred is not really to the point: where the Angoff method can be applied before any response to the items has been collected, the method that is proposed here depends critically on the availability of data.

The judgments of the panel members is an independent source of information, and their judgments will reflect in some sense the quality of the training and their familiarization with the CEFR and with the test material. The section on internal validity in chapter 6 of the manual is relevant here. It is independent (and care should be given to warrant this independence, for example by not allowing panel members to be at the same time members of the construction groups of the test material). As the panels will be very heterogeneously composed, reaching a high degree of inter-judge agreement is highly relevant to the international acceptance (and therefore validity!) of the standards.

An important source of information to be collected independently of the test answers is the judgment of teachers about the level of the sampled students. In principle this information can be asked for each sampled student, but logistic problems may prevent this so that we have to be satisfied with a subsample<sup>4</sup>. The judgment asked from the teachers will not be a holistic judgment as to the CEFR-level of the student, but will consist of a number of can do statements, directly coming from the CEFR or in their DIALANG form. The use of these judgments will be explained more in detail further down. On top of that a number of exemplar tasks will also be used as concretized can do statements, where the teacher is requested to answer whether the students could accomplish these tasks.

In a similar way students will be asked a self assessment using the same can do statements as will be used by the teachers. In principle all sampled students can be asked to fill out a relatively short form stating these self assessment statements. To grant a real independent source of information, these self assessments will be collected well before the testing

---

<sup>4</sup> As students will be sampled randomly within schools, it may happen that the sampled students are instructed by several teachers in the target language, and conversely that some of these teachers will instruct probably only one or two of the sampled students. To avoid large logistic overhead, we might consider to ask judgments only from the two teachers having most sampled students.

proper, although self assessment done closely to the testing seems to lead to higher reliability of the self assessment.<sup>5</sup>

#### **4.5.2 Validation**

The empirical validation itself will consist of comparing the different sources of information with respect to the CEFR levels. As such the procedures can be considered as contributing to convergent validity: the more independent sources one has which result in the same or similar conclusions, the more confident one can be that the conclusions are stable and trustworthy. We treat the different planned methods in turn:

As the test material has been developed in close relation to the CEFR, provisional standards can be derived from the IRT-analysis of the test response data. The standards issuing from the panel judgments can be compared meaningfully to this provisional standards. In a sense this comparison can serve as a validation of the testing framework and the test material derived from it. (This procedure is described in chapter 6 of the manual).

The flexibility of the Cito variation of the bookmark method, described in the previous section allows to use only a subset of the item material in the item map. Depending on the results of the factor analyses, different subsets may be used to set the standards. To avoid as much as possible the effects of differences in leniency between judges, a within judges experimental set up will be used. To make the idea clear, suppose we find essentially two factors which may be identified by the contrast ‘communicative – formal’, then three sets of items will be used, one with predominantly communicative items (i.e., items loading high on the communicative factor), one with predominantly formal items, and one with a mixture representing well the complete test material. Each judge will set standards using the three sets with values of the item parameters resulting from a unidimensional analysis, such that the standards resulting from the three sets are directly comparable. Small differences between the three methods are strong evidence for the construct validity of the standards (and indirectly for the CEFR itself), while large differences are problematic for the linking process.

The can do statements used in the assessment by the teacher and in the self assessments can formally be treated as items, and enter an overall data analysis using IRT, that will at the student level consist of three kinds of items per skill :

1. The items the student answered in the Reading test (say);
2. The assessment by the teacher using can do statements and possibly exemplar tasks for that skill for this student;
3. The self assessment by the student using can do statements for that skill.

Although the parameter estimates of the teacher assessment and the self assessment can be used a posteriori to judge the correspondence between the standards set by the panel and the location of the assessment items (which are directly associated with the CEFR),

---

<sup>5</sup> This appears from not published research by G. Schneider (Fribourg, Switzerland) and the results mentioned were communicated kindly by B. North.

they can also be used during the standard setting procedure to guide and/or correct the judgments. In this sense they will replace impact data, but their role will be similar: they will show the consequences directly in terms of the CEFR of the standards set.

To have a more concrete idea on the procedure, assume that Figure 6.4 (see box) has been used in the first round of the standard setting, and that some judge has set his own preferred standard for A2/B1 as indicated by the vertical line. In a subsequent round (second or third) the judge may see the same item map, but with some 'can do items' added. See Figure 4.3 (below), where one such item has been added as a bold line between the original items 12 and 13.

Suppose this can do statement is a statement at level A2. From the figure, it can be deduced that this judge has set his provisional standard at a level where an A2 can do statement is not even reached at a 50% mastery level, and may conclude from this that his A2/B1 standard has been far too lenient.

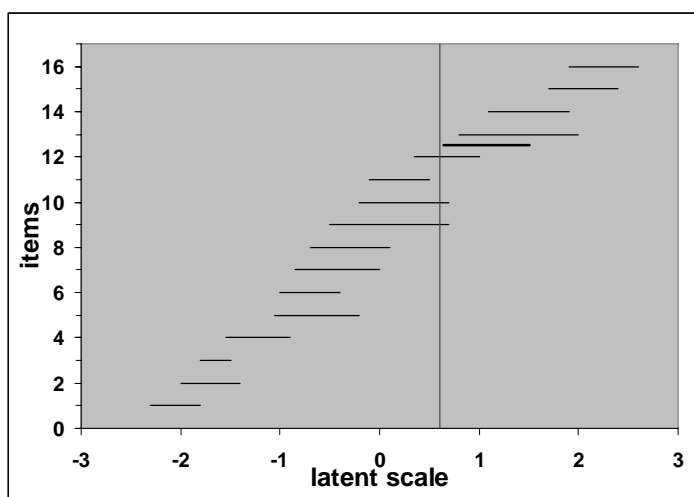


Figure 4.3 Item map enhanced with one 'can do' item

The cross language comparability of the standards can be validated quite directly: by the ranking method (if successful), scales for the same skill and for different languages can be made comparable. Since the standard setting will be carried out by independent panels for each language, the standards (points on the latent scales) are by this procedure automatically brought on the same common scale and their correspondence for different languages can be judged.

# References

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement*, 6, 202-223.

Cizek, G.J. & Bunch, M.B. (2007). *Standard Setting*. Thousand Oaks: Sage.

Council of Europe (2009). *Manual for relating language examinations to the Common European Framework of Reference for languages: learning, teaching, assessment (CEFR)*. Strasbourg: Council of Europe.

Noijons, J. & Kuijper, H. (2006). *Mapping the Dutch foreign language state examinations onto the Common European Framework of Reference*. Arnhem: Cito.

# 5 Standard Setting from a Psychometric Point of View

Gunter Maris, Cito & University of Amsterdam

## 1 Introduction

In this paper an attempt is made to relate some of the problems commonly encountered with standard setting to the psychometric literature. Making these relationships explicit has the advantage that statistical methodology developed for other purposes (e.g., for DIF analyses and person fit) becomes available for use with standard setting.

## 2 Purpose of Standard Setting

The purpose of standard setting is to determine which is the lowest test score with which one is considered to be of level B1 (say). Many procedures are presented in the literature for determining this lowest test score (e.g., Hambleton & Pitoniak, 2006, and further literature cited there).

Ultimately, every standard setting procedure involves two ingredients:

1. Responses of students on the items in the test for which the standard is to be determined (the target test)
2. Responses of those same students on items for which the standard (the so-called golden standard) is already set in a reference test.

This immediately implies that there is a close connection between standard setting and equating methods. Equating methods offer procedures to translate scores from one test to another. The field of equating is well developed and many procedures are available that are well studied and understood (e.g., Kolen & Brennan, 2004; Holland & Dorans, 2006).

Schematically, the data can be depicted as in Figure 5.1. In this figure, the rectangles refer to data tables, where every row corresponds to a student, and every column corresponds to an item. Hence, two rectangles next to each other refer to different tests made by the same students. Two rectangles on top of each other, refer to different students that made the same test. When a rectangle is filled with a question mark, the observations are missing.

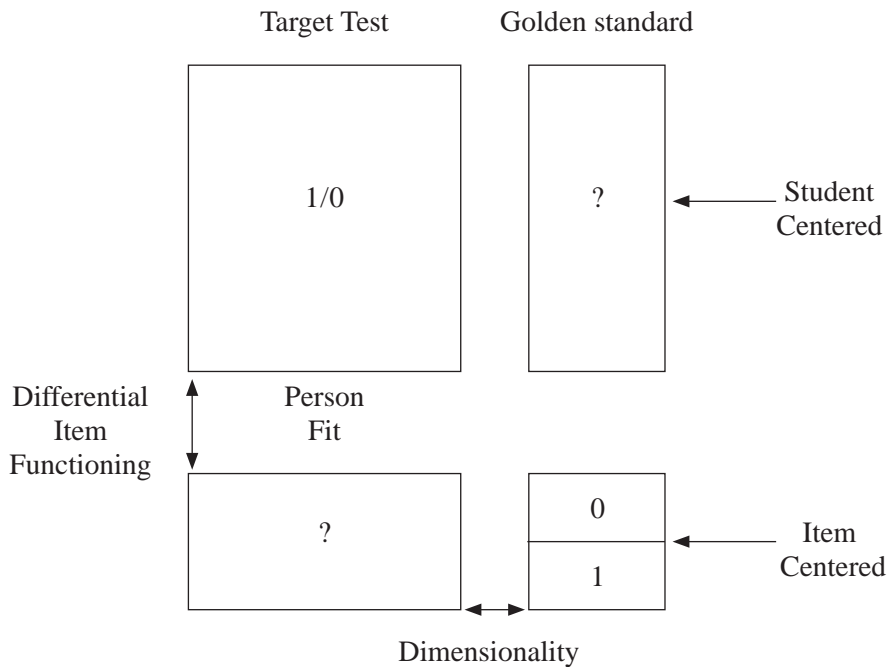


Figure 5.1 A schematic data layout for standard setting

The situation where both the item responses on the target test and the value on the golden standard are observed will be rare in practice. In practice, either the item responses are missing and the value on the golden standard is given; or the value on the golden standard is missing but item responses are given, possibly together with additional information about particular students. Human judges are called upon to supply the missing information. It is in supplying the missing information that the actual standard setting takes place, and this is a purely judgemental problem. After the missing data have been filled in, we only need to transfer the standard to the target test (equating scores), which is a purely methodological problem. In general we can distinguish between two broad classes of standard setting methods depending on which part of the information is missing:

1. Item-centered standard setting: The responses of students on the items in the test for which the standard is to be determined are missing; and
2. Person-centered standard setting: The responses of students on the items in the test for which the standard is already set are missing.

With both types of standard setting, expert judges are used to supply, or impute, the missing responses. For both types of procedures it is important to judge whether the imputed missing data are valid. In the following we will focus on:

1. Equating methods;
2. Validity issues with item-centered standard setting methods;
3. Validity issues with person-centered standard setting methods.

### 3 Equating

The test results of students on a test for which the standard is already set can always be condensed, for the case where a single standard is to be set, to a single binary observation per student, the so-called golden standard. Either the students score is higher than or equal to the standard and the value of the indicator variable is one; or the students score is smaller than the standard and the value of the indicator variable is zero. Depending on the length and other characteristics of the test this indicator variable is more or less reliable, and the reliability will need to be accounted for.

Equating procedures for standard setting can be developed with more or fewer assumptions; either

1. The target and reference test measure the same construct; or
2. The target and reference test measure different, but related, constructs.

The latter case can be treated as a case of predictive validity; whereas the former case can be treated within the framework of item response theory (IRT) (Yen & Fitzpatrick, 2006).

#### 3.1 IRT equating

If the target and reference test measure the same construct, the responses of students on both can be analyzed simultaneously. If a large enough sample of students is used, statistical methodology can be used to evaluate the assumption that both tests measure the same construct (Verhelst, 2001). Through the relation between scores on the reference test and the latent construct we can transfer the standard from observed reference test scores to values of the latent construct; and through the relation between the latent construct and scores on the target test we can transfer the standard from the latent construct to observed scores on the target test (Kolen & Brennan, 2004; Holland & Dorans, 2006).

If the golden standard is reduced to a single binary variable, we effectively estimate its item response function (IRF). The IRF gives the probability with which a person answers an item correct as a function of his/her ability. IRT models differ from each other in the number and kind of item characteristics they contain, and in the functional form of the IRF. Here we consider two item characteristics that are important for our present purposes. First, items can differ in difficulty. The more difficult an item, the lower the probability for giving a correct answer to it, regardless the ability level of a student. Second, items can differ in discrimination. If item discrimination increases, the probability for giving a correct answer will increase for students with ability levels above the item difficulty, but it will decrease for those with ability levels below the item difficulty. The higher the discrimination, the better we can distinguish between students that do and do not meet the golden standard. The effect of both item difficulty and of item discrimination on the IRF is illustrated in Figure 5.2 below.

The difficulty parameter of the golden standard relates to the ability level needed to reach the standard; whereas the discrimination parameter relates to the reliability of the golden standard.

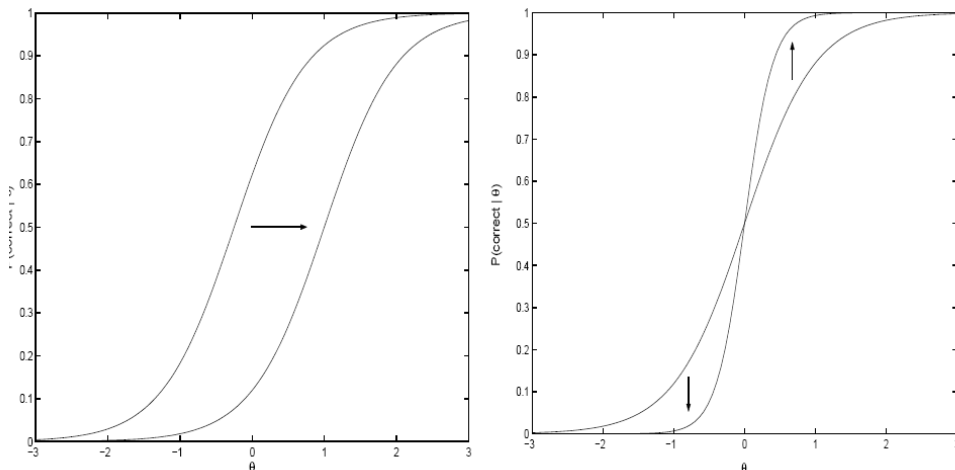


Figure 5.2 The effect of an increase in item difficulty on the IRF (left panel), and of an increase in item discrimination (right panel).

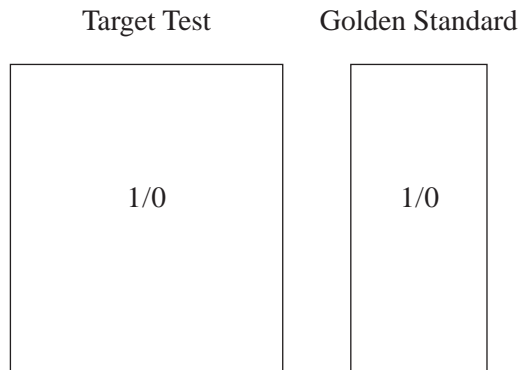
### 3.2 Predictive validity

If the constructs measured by the target and reference tests are not assumed to be identical, but merely related, different methods can be employed. We focus on a situation where there is a true golden standard. This situation is best illustrated with a fictitious medical example. A doctor has to decide whether a patient is suffering from a certain disease, and the best way to determine this is through a post mortem analysis of, say, some internal organ. So, after the patient has died, the diagnosis can be made without error. Clearly, in order to treat the illness, a diagnosis needs to be made before the patient dies. For that reason a medical check list is constructed with questions that are known to be indicative of this particular disease. The problem our doctor now faces is that he has to decide, based on the administration of the check list, when a patient will be classified as being ill and when as being healthy. For that reason an experiment is conducted. A representative sample of patients is administered the medical check list, and after the patient dies the true diagnosis is established. The data from this experiment can be depicted as in Figure 5.3, where the term target test refers to the medical checklist, and the term golden standard refers to the true diagnosis established via post mortem tissue analysis. Having conducted the experiment, the doctor still needs to determine when a patient gets classified as being ill, based on the responses to the medical check list.

If we assume that the medical check list from the previous section contains  $n$  binary questions, there are  $2^n$  different response patterns to consider. For example, with only 10 questions, we need to consider 1024 different response patterns. If the sample of patients



in the experiment is large enough we could in principle for each of the response patterns determine the proportion of patients that actually suffer from the disease. The problem of finding a rule for classifying patients based on their answer to a medical check list is reduced to choosing a cut-off value for the probability of being truly ill given the response pattern of a patient. It is important to observe that the choice of a cut-off value for the probability of being truly ill, given the fact that the response pattern is essentially arbitrary. In practice, the situation will rarely be this simple.



*Figure 5.3 A schematic data layout for standard setting*

#### 4 Item-centered methods

With an item-centered standard setting procedure, expert judges are asked to supply the item responses of a candidate given the value of the golden standard. In order to validate the imputation procedure there are two important aspects:

1. The characteristics of the items as based on real responses and based on imputed responses should be the same. That is, there should be no Differential Item Functioning (DIF) between expert judges and real students.
2. The pattern of item responses produced by an expert judge should be one that could have been produced by a real student. A person fit analysis, with the expert judge as person, can be used to evaluate if a response pattern could have been produced by a real student.

The first aspect concerns the behaviour of the pool of expert judges; whereas the second aspect concerns the behaviour of individual judges.

An item is said to show DIF if its operating characteristics (e.g., item difficulty and/or item discrimination) are different for different populations. In the context of standard setting methods, the relevant populations are those of real students and those of judges. If according to the pseudo item responses supplied by human judges, one item is more difficult compared to another; whereas for real students the reverse holds, the judges, as a group, did not succeed in performing their task.

One reason why items show DIF between judges and real students is that one or two of the judges show systematic deviations in their behaviour. A person fit analysis can reveal such

deviations. In principle, a person fit analysis reveals whether a particular response pattern produced by a judge could have been the response pattern produced by a real student. If the response pattern of a judge is very unlikely to have been the response pattern of a real student, we conclude that this particular judge did not succeed in performing his/her task. Observe that we do not expect different judges to produce the exact same response patterns, even though they are instructed to consider the same imaginary student. This poses no real problems because even students with the exact same ability level are not expected to answer the exact same items correct.

However, if all judges consider the same imaginary student, the amount of variation in the scores they produced should be consistent with a single ability level. Whether or not this is the case can, and should, be evaluated explicitly. Specifically, an IRT model specifies not only an IRF for every item, but from these IRFs we can derive a score distribution for every ability level. If the score distribution produced by the judges is consistent with a particular ability level, then this level automatically is the golden standard. If this is not the case, the judges disagree, to some extent, about the exact position of the golden standard.

## **5 Student centered methods**

With a student-centered standard setting procedure, expert judges are asked to supply the item responses on the reference test for a real student that took the target test. That is, the judges are relied on to tell us whether or not the students meet the golden standard.

In order to validate the imputation procedure we need to evaluate the item characteristics of the golden standard. For instance, if the relation between the golden standard and the score on the target test is not monotone, it will be difficult, if not impossible, to equate the golden standard to the scores on the target test.

Additionally, if each judge considers a sufficiently large number of students, we may use DIF analyses to evaluate whether the IRF of the golden standard differs between judges. If DIF occurs between judges we have clear evidence that not all judges employ the same standard. In its most basic form, we estimate the IRF of the golden standard, or equivalently the ability distribution for those who do, and those who do not meet the golden standard. In practice it will never be the case that there is a single ability level, below which every student fails to meet the golden standard, and above which every student meets the golden standard. The item discrimination parameter of the IRF of the golden standard provides information about the reliability of the golden standard. There are however no clear cut criteria for determining whether or not the discrimination parameters are large enough. This in contrast to the situation with item-centered standard setting methods where we can statistically test, in principle, whether the judges all used the same ability level in their judgements.

## **6 Conclusion**

In this paper the process of standard setting was cast in a psychometric framework to reveal that many of the standard psychometric problems (e.g., dimensionality, DIF, etc.) also occur with standard setting. It would therefore seem advisable to consult the psychometric literature referred to above on issues that relate to standard setting, to avoid some of the pitfalls discussed in this paper.

# References

Hambleton, R. H., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement*. Westport: ACE/Praeger series on higher education.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement*. Westport: ACE/Praeger series on higher education.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking. methods and practices* (Second ed.). New York: Springer.

Verhelst, N. (2001). Testing the unidimensionality assumption of the rasch model. *Methods of Psychological Research Online*, 6 (3), 231-271.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement*. Westport: ACE/Praeger series on higher education.



Part II  
Accounts from practice:  
reports from the practitioners



## 6 Relating the Trinity College London International ESOL examinations to the CEFR

Cathy Taylor, Trinity College London

### Introduction

This report<sup>1</sup> will summarise the CEFR calibration project which Trinity undertook 2005-2006, led by Spiros Papageorgiou, Lancaster University. The aim of the calibration project was to link the Graded Examinations in Spoken English (GESE) and the Integrated Skills in English (ISE) suites of language examinations to the CEFR using the pilot linking manual. However, due to the brevity of this report only the GESE mapping will be discussed. The report will cover the stages of the project, provide samples of the methods used and results, as well as consider some of the problems encountered in the mapping process. The action taken by TCL after the project and the programme for future research will be discussed.

### Purpose of the project

Following the publication of the *Common European Framework of Reference for Language: Learning, Teaching, Assessment* (Council of Europe, 2001) it became apparent that language testing now had a common reference point and that transparency and comparability between language tests would be possible. The Council of Europe duly published a pilot version of the Manual for Relating Language Examinations to *Common European Framework of Reference for Language: Learning, Teaching, Assessment* (Council of Europe, 2003) and invited exam providers to pilot it and provide feedback on the linking process. Trinity College welcomed the opportunity to take part in the piloting of the manual whilst mapping and standardising the TCL International ESOL examinations to the CEFR. TCL commissioned Spiros Papageorgiou from Lancaster University to lead the project. The project took place over 18 months from 2005-2006. The final report was published in February 2007.

### Outline of TCL ESOL examinations

The International ESOL examinations mapped to the CEFR are the Graded Examinations in Spoken English (GESE) and the Integrated Skills in English (ISE). The former are oral examinations based on a 1-1, face-to-face interview with an examiner. There are 12 levels, equally divided into 4 stages, ranging from beginners to full mastery. Results attainable are Pass, Merit or Distinction. The latter assesses reading, writing, speaking and listening in an integrated fashion comprising of a portfolio, traditional controlled written and oral components. At the time of the project ISE comprised of 4 levels.

<sup>1</sup> The full report is available from <http://www.trinitycollege.co.uk/site/search.php#webpages>

## Methodology

The linking project comprised of four sets of interrelated activities:

- 1 **Familiarisation.** This stage, which was repeated before Specification and Standardisation, was imperative in order to ensure that the members of the linking panel were familiar with the content of the CEFR and its scales.
  - 2 **Specification.** This stage involved the description of the content of the test to be related to the CEFR first on its own right and then in relation to the levels and categories of the CEFR.
  - 3 **Standardisation.** The outcome of this stage is the reinforcement of the previous claim. Standardisation involves achieving a common understanding of the CEFR levels illustrated by examples of actual learners' performance.
  - 4 **Empirical validation.** In this project Trintiy carried out an internal validation study. This phase will not be discussed in this report due to space limitations.
- The tasks and pro forma suggested and supplied by the Manual were used in all the activities above.

### The panel of judges

A panel of 12 judges was selected. All the members are involved in with the examinations in some capacity, e.g: test design, item writing, quality assurance and validation, senior examiners, plus two members of the Head Office Academic Team. The group were to some extent already familiar with the CEFR. Although all the judges are familiar with the exams it is important to note that, apart from the members of the Academic team, all are free lance and work in other areas of ESOL.

### Familiarisation

Before the session the judges were asked to familiarise themselves with the scaled descriptors. In the session judges were given a booklet containing samples of the scaled descriptors from Tables 6.1, 6.2 & 6.3, global, skills and qualities descriptors respectively, of the CEFR. These were broken down into their constituent parts and judges assigned a CEFR level to the descriptor. A further task using Table 5.8, written assessment criteria grid, required judges to put the cut-up descriptor into the correct cell. The results were displayed on excel and a discussion followed. The judges found this helpful in clarifying slight differences between the descriptors. A paired discussion followed the confetti-style allocation task for Table 5.8. The results were analysed for inter-rater reliability and inter-rater reliability and the CEFR.

Table 6.1 presents the agreement between the panellists' level assignment with the correct level. Spearman correlations were run between each judge's levels and the correct CEFR levels. These correlations are all very high. It should be stressed here that Spearman coefficient shows rank order correlations, which in this context means that it explains agreement in the order that two sets of descriptors had been arranged and should not be interpreted as exact agreement of assigned levels. Even a correlation of 1 can occur with 0% exact agreement if different ranges of the scale are used as pointed out by Kaftandjieva (2004:24); for this reason, another coefficient is included: Cohen's (Kappa) calculates exact



agreement by also taking into account agreement by chance, which cannot be taken into account when reporting raw scores. Kappa is reported along with Spearman correlations below. In the Kaftandjieva and Takala (2002) study coefficients above .7 are reported as satisfactory.

*Table 6.1 Rater-CEFR agreement-summary statistics*

Scales	Spearman correlations			Cohen's Kappa			
	Mean*	Min	Max	Mean	Min	Max	N
Speaking1	0.911	0.871	0.928	0.464	0.28	0.602	10
Speaking 2	0.958	0.913	0.985	0.626	0.282	0.88	12
Writing 1	0.883	0.791	0.938	0.423	0.228	0.516	11
Writing 2	0.907	0.828	0.957	0.547	0.335	0.709	10
Listening 1	0.907	0.832	0.961	0.548	0.408	0.74	11
Listening 2	0.920	0.855	0.962	0.593	0.422	0.805	12
Reading 1	0.959	0.901	1	0.591	0.235	1	11
Reading 2	0.968	0.923	0.994	0.687	0.474	0.939	12
Global 1	0.939	0.901	1.000	0.589	0.36	0.84	12
Global 2	0.959	0.923	0.994	0.66	0.439	0.88	12

\*Average using Fisher's Z-transformation

### **Specification**

Before the meeting forms A1-A7 were completed by the Academic Team in Head Office. In the subsequent session forms A8, A11, A13, A9 A19-21 were chosen for GESE and ISE. Forms A10, A12, A14 - A16 were also used for ISE. The syllabuses for each suite were used along with the CEFR. The judges worked in groups with 3 GESE grades and 2 ISE levels given the amount of form filling required for this stage. The judges experienced difficulties with terminology at this stage and lengthy plenary sessions were needed to establish common understanding of the terminology. Once the process had begun it gathered momentum as the group became more familiar with the task. However, the judges found that there was a considerable amount of repetition in the form filling.

Given the number of levels in GESE the branching approach (Council of Europe, 2001:31-33) was found eminently suitable. However, not all the GESE grades were a good fit within the branching approach, with only parts of a grade fitting part of a descriptor. The judges had difficulty in assigning GESE grades to the categories in Table 5 *External context of use: descriptive categories* since in the Topic phase of GESE candidates talk about a topic of their choice. Further problems were encountered in that the descriptors were the same for C1 and C2.

### Standardisation and benchmarking

Following a training and benchmarking (CoE DVD and listening files from DIALANG) judges were shown samples of GESE grades and using a minimum of 3 CEFR scales assigned a CEFR level to the grade. However, as the judges progressed through the grades up to 17 scales were used. Due to time constraints in the project only 1 sample from each grade was used. In future studies we would like to use more samples in assigning CEFR levels.

Table 6.2 summarises coefficients of consistency and agreement. All judges watched DVD samples from Trinity Grades and provided ratings of the candidates' performance using the CEFR scales. Group discussion followed the rating of the Grades of each GESE stage. In general internal consistency of judgements and agreement were very high. Indices were calculated by converting judges' ratings using Fout! Verwijzingsbron niet gevonden. in the CEFR report p 43 and then using SPSS; this applies to all sets of ratings in the Benchmarking sessions.

*Table 6.2 Agreement and consistency of judges-GESE benchmarking*

Stage	Inter-rater reliability			Alpha	ICC**	W**
	Mean*	Min	Max			
Benchmarking	0.954	0.888	0.983	0.994	.932	.978***

\* Average using Fisher's Z-transformation

\*\* Statistically significant at level pff.01

\*\*\*Only overall spoken interaction scale

### Establishing cut-off scores

In this phase the judges provided the estimated level for candidates receiving each of the scores in the GESE suite in the first round and in the second round they only estimated the CEFR level of the borderline candidate for each GESE grade.

### Results

Table 6.3 Represents the CEFR level of borderline and secure pass candidates in the GESE suite.

Table 6.3 CEFR level of borderline and secure pass candidates in the GESE suite

Grades	Level of borderline candidate	Level of secure pass candidate
Grade 12	C1+	C2
Grade 11	C1	C1
Grade 10	B2+	C1
Grade 9	B2+	B2
Grade 8	B2	B2
Grade 7	B1+	B2
Grade 6	B1	B1
Grade 5	B1	B1
Grade 4	A2+	A2
Grade 3	A2	A2
Grade 2	A1	A1
Grade 1	Below A1	A1

**Action taken by Trinity College following the CEFR mapping project**

In 2007 the GESE and ISE syllabuses were due for revision. In the light of the findings of the project it was deemed appropriate to alter the syllabus to be more in line with the CEFR and the following significant revisions were made:

- Grade 1 was not allocated a CEFR level since it was considered below the A1 descriptor.
- Grades 2, 7, 8 and 12 the requirements were revised to reflect CEFR competences.

TCL was very happy to take part in this project and appraise the International examinations from a different perspective. The empirical studies have not been mentioned in this report but TCL plans to undertake further analyses, both internal and external. Participating in this project is a springboard for future research and calibration projects.



## 7 Analyzing the decision-making process of standard setting participants

Spiros Papageorgiou, University of Michigan

Investigations of the thought processes and experiences of judges in non CEFR-related situations have been conducted as part of validating the outcome of a standard setting meeting (cf. McGinty, 2005). Such an investigation into the judges' thought processes and experiences employed qualitative methods (Buckendahl, 2005) but has not been attempted in the CEFR linking context to date. This study attempts to fill in this gap. It explores the factors that affect the judges' decision-making during the Standardization stage presented in the Council of Europe's Manual and the problems the judges faced during this stage. The analysis showed that decision-making was affected by a number of factors that were irrelevant to the judgement task and that the standard setting process was not without problems for the judges. The implications for examination providers wishing to standardize their tests on the CEFR are considered.

### Introduction

Investigating judgments during the linking process is fundamental for the validity of the CEFR linkage. This is because the CEFR linking process, having standard setting in its core, depends on human judgment (Kaftandjieva, 2004: 4). Because of this dependence on human judgment, standard setting specialists in the US have started investigating the thought processes and experiences of judges in non CEFR-related situations as part of validating the outcome of a standard setting meeting (cf. McGinty, 2005). Such an investigation into the thought processes and experiences of judges employed qualitative methods (Buckendahl, 2005) and has not been attempted in the CEFR linking context to date. Consequently, this part of the validity of the CEFR linkage is still unexplored.

The present study attempts to bridge this gap in the CEFR literature by exploring the factors that affect the judges' decision-making during the Standardization stage of the CEFR linking process and the problems the judges faced during this process, as they are very likely to impact on the linking claim that is produced after the completion of the Manual's (Council of Europe, 2003) Standardization stage.

### The study

The study was organized as part of the Trinity College London CEFR project (Papageorgiou, 2007), also described by Taylor in this volume. The project aimed to relate two examinations to the CEFR, the Graded Examination in Spoken English (GESE) and Integrated Skills in English examination (ISE). The data were collected over three days in February/March 2006, during the Standardization stage presented in the Manual with the

participation of 11 judges.

The study addressed two research questions (RQs).

*RQ1: Which factors affect decision-making during the Standardization stage?*

*RQ2: What problems do the judges face when judging performance and setting cut-off scores in relation to the CEFR?*

## **Methodology**

11 judges, who were involved in various stages of the test development process of the two exams, participated in the project. After conducting their individual rating tasks, the judges were invited to discuss their experience with the rest of the panel and explain how they made decisions. Approximately 2 hours of group discussions were recorded. The recordings were fully transcribed and imported into Atlas.ti (Muhr, 2004) a program that provides help in systematically coding qualitative data. After importing the data into Atlas.ti, the general sense of the data was explored by reading them several times and a coding frame was then created that would help organize findings and report results. The coding frame was based both on both deductive (i.e. drawing codes from existing theory) and inductive (i.e. drawing codes from the raw data) approaches. Deductive coding was employed to address RQ 1 (factors affecting decision-making) and it was based on standard setting research presented in Buckendahl (2005). Inductive coding addressed RQ2 and aimed to provide a systematic presentation of possible problems the judges faced when using the CEFR and its descriptors to judge examinee performance and set cut-off scores. A second researcher coded a sample of the data in order to examine the reliability of the coding frame and to refine it.

## **Results**

Analysis of the group discussion data with regard to RQ1 (factors affecting decision making) showed that the judges mainly made their decisions after carefully examining the CEFR descriptors. However, other factors also came into this judgment-making process, namely: judging the Council of Europe samples based on candidates the judges had previously examined, expectations of the CEFR level of the candidates due to insiders' bias, the notion of the borderline examinee, the exam's rating scale and finally, reporting compensatory composite scores. The last point is illustrated in the extract below. A judge comments on the compensatory composite score awarded by the exam provider because a candidate would be awarded an overall pass if he/she scored three Cs (pass) and two Ds (fail) out of five components in the exam. Because of failing two components, the judges decided to lower the initially intended level (C2).

*Tim: a two Ds and three Cs profile means. I don't think it is arguable. if it means they are C just about C2 at three points and not C2 at two points. and we are saying they are not C2 all the time (Standard setting discussion after round 1, P11: 163-170)*

As is the case in relevant research on the CEFR in other contexts (e.g. Alderson et al., 2006) the judges reported some problems which hindered the judgement task (RQ2), namely:

1. the context-free character of the CEFR;
2. the description of real-life language use in the CEFR scales;
3. the wording of the CEFR descriptors;
4. the unsuitability of the CEFR scales for judging young learners;
5. the quality of some Council of Europe samples;
6. the lack of descriptors for particular aspects of performance and levels;
7. their own difficulty grasping the notion of the borderline examinee.

The second point is illustrated in the extract below. Since the CEFR scales are behavioural scales, they attempt to describe language in real-life contexts. However, the judges found that such real-life descriptions were not helpful when making judgements in an exam context for example regarding the notion of repair (A1 Interaction scale, CEFR Table 3) because of this real-life description.

*Alice: I mean the descriptors describe real life situation where you have an exchange of information and also things break down a lot so there is got to be more repair. there is very little repair over here. because the candidates were prepared and the exam is very structured*

*Tim: this comes back to what we were talking about yesterday doesn't it? because the descriptors are not written for use with an examination*

*Alice: exactly they are written to describe a situation in real life (Benchmarking GESE Initial, P4: 36-46)*

## Conclusion

The findings of the qualitative analysis during the Standardization stage have important repercussions for claiming linkage to the CEFR. The judgment of examinee performance might be affected by factors that are irrelevant to the CEFR scales. Such factors should be controlled during the standard setting meeting, as the aim of the Standardization stage is to have substantive judgments of examinee performance based on the standard (i.e. the CEFR scales), not on the judges' personal understanding of the CEFR scales or their prior knowledge as to the level of candidates taking the exam.

Moreover, the effect of the exam's rating scale and compensatory scoring on judgments during the Standardization stage leads to two implications for examination providers involved in the CEFR linking process. The first implication is that rating scales, like test specifications, should be based on the CEFR. The second implication is that compensatory scoring can be misleading when reporting results in relation to the CEFR, as someone passing the exam might be wrongly considered to be able to perform at a specific CEFR level, even without having passed one or more components of the exam.

It should be stressed that the above points are important because of the consequences of a linking claim on various test users. Uses of tests and interpretations of results might be based on the CEFR level claimed by a test provider, for example when teachers choose

examinations for their students, when universities admit students and when employers recruit new staff. Inevitably, invalid claims as to the CEFR level of a test might affect the appropriateness of such decisions.



# References

- Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & 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.
- Buckendahl, C. W. (2005). Guest editor's introduction: Qualitative inquiries of participants' experiences with standard setting. *Applied Measurement in Education*, 18(3), 219-221.
- Council of Europe. (2003). *Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. Preliminary pilot version*. Strasbourg: Council of Europe.
- Kaftandjieva, F. (2004). *Standard setting. Section B of the Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe.
- McGinty, D. (2005). Illuminating the “Black Box” of standard setting: An exploratory qualitative study. *Applied Measurement in Education*, 18(3), 269-287.
- Muhr, T. (2004). Atlas.ti (Version 5.2). Berlin: ATLAS.ti Scientific Software Development GmbH.
- Papageorgiou, S. (2007). *Relating the Trinity College London GESE and ISE exams to the Common European Framework of Reference: Piloting of the Council of Europe draft Manual*. (Final project report). Lancaster: Lancaster University.



## 8 Benchmarking of videotaped oral performances in terms of the CEFR

Gunter Maris, José Noijons and Evelyn Reichard, Cito

### Summary

This paper describes an international activity that has focussed on the benchmarking of speaking performances based on the descriptors in the CEFR. Cito and the Dutch Institute for Curriculum Development (SLO) have collected video performances to illustrate CEFR levels in spoken interaction. In the benchmarking process it was found that judges have had difficulty in classifying individual performances in the finer grained CEFR categories (A1, A2, B1, B2, C1 and C2). If a coarser classification is used, such as a distinction in levels A, B, and C, agreement improved considerably and acceptable levels of agreement could be reached for most performances included in this study.

### Introduction

Although the CEFR is based on descriptions of language behaviour, until recently stakeholders sometimes had difficulty in understanding what exactly was meant by a particular descriptor (*can-do statement*). Descriptors are sometimes phrased in linguistic jargon not accessible to learners. Also, descriptors may differ in the detail necessary to understand how well a language user has to perform to be at a particular level. The video materials that have been developed may be of use to learners and teachers in the learning and teaching process. This activity is an on-going process that has started in 2007.

### Method

#### Production of tasks and videotaped performances

As the samples of speaking performances were to serve a didactic purpose and the greater part of Dutch learners and speakers of English is judged to be in the A2-B2 range, the majority of tasks was produced for this range. However, some tasks were also developed on the basis of *can-do* statements at A1 level and in the C1-C2 range. Some of the contexts in these tasks are aimed at adult learners and speakers and other contexts are aimed at young adolescents. All tasks have been developed for oral interaction between two interlocutors, of whom one is a native speaker of English and the other has the role of a non-native speaker.

Participants were hired from a student population in the age group 12–18 and from an adult population. Care was taken that in both groups (students and adults) all (pre-judged) CEFR levels were represented. All participants allowed Cito and SLO to use their videotaped performances for educational purposes. It was made clear to participants that the aim of

the project was not to judge persons, but to collect performances that could serve as examples of what was understood by descriptors in the CEFR. Not only were the performances gathered in a non-testing environment, the tasks themselves were produced independent of any existing test-format. In this way it was hoped that the benchmarks could be used in multiple learning and testing environments.

A sample was made of performances to be used in the benchmarking procedure. This sample includes performances that were pre-estimated to be *below* the level required by the task, *on* this level and *above* this level. Three CD-ROMs have been produced, each containing a different sample of 20 video-taped performances. A number of performances have been used as anchors: they appear on more than one CD-ROM.

### **Instruction of judges**

For the benchmarking procedure it was important to select knowledgeable judges who are active in the field of second language learning and testing. These have been recruited from among members of the ALTE and EALTA organisations. Prospective judges were asked to read tests taken from the CEFR and carry out familiarisation exercises as described in the Manual for relating language examinations to the CEFR, published by the Council of Europe. Cito and SLO have had no control over whether these instructions have been carried out as described in the Manual (Council of Europe 2003). However, there are a number of reasons why there is some confidence that procedures have been followed. First, the judges are staff members of reputable testing and teaching organisations, second their participation was voluntary and third there was general agreement with the goals of the study and their role in achieving these goals. Also Cito had much experience in instructing judges from other linking activities (Noijons, 2006).

### **Judgment procedure**

The benchmarking procedure has focused on two judgements: (1) *what level is the task at?* and (2) *what level is the performance at?* In the benchmarking procedure used here, judges first had to assign levels to the tasks. After that judges were asked to rate a performance on that task using the criteria (taken from the CEFR) for the level that they judged the task was at. If speakers do *not* perform at the relevant level, they may be *underperforming* or *overperforming*. If they are *underperforming* we may have to give them a task at a lower level and see how they perform at that level. We can *not* say that if the speaker does not perform at the level of the task he is performing at the next lower level. The same goes with *overperforming* students: we may have to give them a task at a higher level to see how they perform at that level. We can *not* say that if speakers perform better than required at the level of the task, they are performing at the next higher level. We have therefore instructed judges to indicate if speakers performed *at* the level of the task, *under* the level of the task or *above* the level of the task.

## Results and discussion

### Judging the task level

In the benchmarking process two steps have been distinguished: judging the task level (step 1) and judging the performance level (step 2). Judges were first instructed to determine the level of the task. The level that was *most often* chosen by judges for each task was termed the *modal level*. We found that there was no overwhelming consensus among the judges regarding the level of many tasks. If we considered a broad classification of levels and only distinguished between levels A, B, and C, we found that a more acceptable level of agreement existed between judges. The broader classification in three levels (A-B-C) did not suffice for all tasks. Sometimes a range such as A2-B1 was found to be the modal level.

### Judging the performance level

We have instructed judges to indicate if speakers performed at the level of the task, under the level of the task or above the level of the task. When a judge deems a performance to be above (below) the level of the task this is interpreted as meaning that the task level as well as lower (higher) levels of performance are excluded from consideration. Put differently, if for a task judged to be at the B1 level a performance is considered to be above task level the judge is expected to agree to the following three statements:

- This cannot be the performance of an A1 candidate.
- This cannot be the performance of an A2 candidate.
- This cannot be the performance of a B1 candidate.

Hence, based on the judgements, performance levels are excluded from consideration. When a judge deems a performance to be at the level of the task, all other levels are automatically excluded from consideration.

Ideally we find that across judges there is a single level which is never excluded from consideration. In such a situation we may safely conclude that this is the level appropriate for this performance. In reality we cannot expect this ideal to be realized. We consider two approaches for determining the level appropriate for a performance. First, we may look for a single level which is hardly ever excluded from consideration. For every level we can determine what percentage of the judges excludes it from consideration, and that level for which this percentage is lowest would be the most appropriate level. This level we call the modal level. It is of course arbitrary which precise level of agreement between judges is deemed sufficient. Notice that the modal level need not be unique (i.e., there may be multiple levels with the same exclusion percentages).

If there is no single level that appropriately describes a particular performance, we may look for a range of levels that would adequately fit a particular performance. To the modal level we may add the next higher or the next lower level. To choose between the two, we look at the percentage of judges which excluded the modal level but included the next lower or the next higher level. If there is a larger percentage of judges that excluded the modal level but did not exclude the next lower level compared to the percentage of judges

that excluded the modal level but did not exclude the next higher level, we have added the next lower level to the modal level to obtain an appropriate performance range. Obviously, for a range of levels we also needed to determine whether it was appropriate for this performance in much the same way as for the modal level. If a large percentage of the judges still excluded both levels the range was not appropriate and another level may have needed to be added.

Results have been produced in tabular form. These tables give, for each aspect, exclusion percentages for either the modal level or the modal range. It has been found that in general exclusion percentages are considerable for modal levels but much more acceptable for modal ranges.

### **Conclusions and topics for further study**

Possibly the most striking conclusion that can be drawn from this benchmarking exercise is that judges have had difficulty in classifying individual performances in the finer grained CEFR categories (A1, A2, B1, B2, C1 and C2). If a coarser classification is used, such as a distinction between levels A, B, and C, the situation improves considerably and acceptable levels of agreement can be reached for most performances included in this study. However, this is not how the developers and the users of the CEFR have been looking at levels. In fact, rather than aggregating levels or referring to ranges, they have come up with suggestions to further subdivide levels, such as in B1-1 and B1-2, or A2 and A2+. From this study it would seem such subdivisions need more validation.

An interesting issue that needs further study is whether across multiple performances of a single candidate a finer grained classification becomes possible. Indeed, judges have also been asked to give so-called comprehensive judgements of a series of performances in the same task level range. It must be remembered though that when benchmarking performances in this comprehensive way, the focus in the benchmarking will be more on a person performing at a level rather than on the performance of a task (the latter having been the intention of the project).

From a test construction point of view, this study has also yielded an interesting finding, even though for the benchmarking per se it was of less relevance. This is that task developers have pre-estimated the tasks they developed at lower levels than the judges have benchmarked the tasks. In the construction of tasks, depending on the status of a test, it is important to have tasks field-trialled. This study shows that when developing oral tasks it is important to have a panel judge the level (difficulty) of a task before such tasks are administered to learners in a test environment.

# References

Council of Europe (2003). *Manual for Relating language examinations to the CEFR; Preliminary Pilot Version*. Strasbourg: Council of Europe

Noijons, José & Henk Kuijper (2006). *Mapping the Dutch Foreign Language State Examinations onto the Common European Framework*. Arnhem: Cito





# 9 Designing Proficiency Levels for English for Primary and Secondary School Students and the Impact of the CEFR

Karmen Pižorn, University of Ljubljana, Faculty of Education, Slovenia

## A Slovenian Experience

### Introduction/Background

In the mid-1990's the education authorities decided to implement school-leaving examinations in all school subjects. The implementation of external national assessment was, therefore, supposed to assist the realisation of the quality principle. Surprisingly, new curriculum documents, which traditionally define what students are to be taught were developed later, i.e. from 1995 to 1998. The foreign language curriculum construct shifted to a more explicit Communicative Approach. It followed the main paradigms established in the Threshold Level (Van Ek and Alexander 1975) and later Threshold 1990 (Van Ek and Trim 1998) and took into consideration the results the students achieved in the external national assessments. However, the foreign language curricula did not detail the language skills that students were expected to develop at different grade levels and were therefore not helping teachers with classroom assessment.

In 2007, the Ministry of Education established a schedule for a curriculum review, which resulted in a two-year review process to ensure that the curricula are kept current, relevant and age-appropriate. By developing and publishing revised curriculum documents for use by all Slovenian teachers and other stakeholders the Ministry of Education of the Republic of Slovenia wanted to set standards for the entire state. The National English Language Curricula Reform Team<sup>1</sup> was set a task of revising two documents, English Language Curriculum for Primary School (pupils aged 9 -14) and English Language Curriculum for Secondary Grammar School (students aged 15 to 18). We<sup>2</sup> revised the existing two curriculum documents in the light of most of the questions defined by Richards (2001), such as whether the curriculum is achieving its goals, how it is being implemented, whether those affected by the curriculum are satisfied with it etc. However, the main task was to develop language proficiency levels for the four language skills for primary and

<sup>1</sup> From now on the Team (consisting of primary and secondary-school language teachers, language advisors, university English language methodologists and others).

<sup>2</sup> From now on in this article 'we' refers to The National English Language Curricula Reform Team.

secondary-school pupils and align the developed levels to the CEFR<sup>3</sup> reference levels. The latter process is also the main topic discussed in this article.

## The Overview of the curriculum revision process

First, we studied Chapter 3 *Criteria for descriptors for Common Reference Levels of the CEFR* and promoted a common understanding of the main terms, especially a scale of reference levels and a level of proficiency. Then we studied the four criteria for developing a scale of reference levels: a) context-freedom, b) context-relevance, c) relevance in terms of language competence theories and d) user-friendliness. The first criterion means that the reference levels must not be produced for one or a certain number of school/s or for a certain student from a certain town or village. The levels also have to take into account the context they are developed for, in our case primary and secondary school students, but be transferable to any primary school pupils in any country in the world. As for the third criterion, we accepted the theory of language as communication (Halliday, 1970; Hymes, 1972; Wilkins, 1976; Canale and Swain, 1980; Brumfit and Johnson, 1979; Savignon, 1983) and Competency-Based Language Teaching (Schneck, 1978; Grognet and Crandall, 1992; Richards, 2001). This approach moves away from the content and process of teaching and shifts the focus to the ends of learning rather than the means. The last criterion, user-friendliness, means that reference levels have to be written in the way that they are easily comprehensible by practitioners (Richards, 2001). We satisfied this criterion by making reference levels available on the website and in a virtual classroom for a few months with an invitation to the teachers to comment on them. The levels were also discussed with the teachers during in-service teacher training workshops run by the Team members.

We also tried to follow the two criteria that refer to measurement issues. Firstly, we took into consideration the data available from the national assessment and combined these with the judges' assessment of language tasks. The second criterion refers to the number of levels, which must show progression in different sectors and be able to show reasonably consistent distinctions. This resulted in four different scales of reference levels for age groups 11, 14, 16 and 18. We concluded that it was not feasible to set national reference levels after/for each school year due, for instance, to the different learning styles and learning abilities of the students.

In May 2007, a three-day workshop on Setting up standards in the language curricula and relating them to the CEFR was held for all National Language Curricula Reform Teams curriculum.<sup>4</sup> A number of topics were discussed, such as overall educational goals, different

---

<sup>3</sup> The CEFR stands for The Common European Framework of reference for Languages: Learning, Teaching and Assessment. It "provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks etc. across Europe". (CEFR, p. 1) In November 2001 a European Union Council Resolution recommended using the CEFR to set up systems of validation of language ability. The six reference levels (A1 to C2) are becoming widely accepted as the standard for grading an individual's language proficiency.

<sup>4</sup> The workshop was led by Prof. Sauli Takala from Finland who has extensive experience with curriculum development and the use of the CEFR in curriculum construction and assessment.

types of language curricula, what a curriculum should include, why it needs to be multidimensional, why it should be transparent to all parties, adaptation or adoption of the CEFR etc. This workshop was of vital importance for as it offered insight into the CEFR as a reference tool. We discovered that there is much more in the CEFR than just proficiency levels.

From June 2007 to November 2007, the members of the Team had to do a number of tasks at home while being supported via email. Each member received the following documents to be studied: global and analytical proficiency scales from the CEFR, background information on descriptors, Slovene Language Portfolios model, English language scales from Finland, the foreign language standards from the USA etc. Each member received 5 charts which they had to fill in. Chart 1 included four language skills and had to be filled in with the information about texts, tasks, length, and topics and any extra information that referred to the age group the member was teaching. This chart was considered to be more general and to give an overview of the four language skills (See Appendix 1a and 1b). Charts 2 to 5 (See Appendix 2a and 2b) referred to listening, reading, speaking and writing language skills and included a number of dimensions that influence language comprehension and/or production. For reading and listening skills, we heavily depended on the Dutch CEF Grid, which we found very useful for taking into account the horizontal level of language learning and proficiency.

In November 2007, a three-day workshop took place and it was intended to design language proficiency levels for all age groups. First of all, the Team members compared their own charts and designed an overall chart including a number of dimensions (see Appendices 3a, 3b and 3c). After finalising the chart, the Team members had some time to revisit the CEFR common reference levels, including DIALANG self-assessment scales. Then all members were assessed individually as to how well they understand the level descriptors and whether they can correctly relate them to the CEFR levels from A1 to C2 (see Appendix 4). The next step led the Team to look at and assess a selected sample of the calibrated tasks/items, e.g. ALTE, DIALANG, EUROCENTRES/MIGROS and the Dutch CEF Grid sample tasks.

As can be seen from Table 9.1, most of the Slovene experts rated DIALANG items higher than DIALANG judges. After the rating process, the Team discussed possible reasons which may be summarised as following:

- The Slovene judges were used to rating reading tasks and their items as a whole. They found the rating of individual discrete items without being able to read the whole text quite confusing.
- The wording in multiple choice items was sometimes felt to be more difficult than the input text itself.
- Some of the judges were uneasy about the multiple-choice test method.
- Some judges believed that the tasks would have been easier if the task instructions, e.g. for multiple-choice, had been in the students' mother tongue.

Table 9.1 English DIALANG Reading items as rated by ten judges

Test item	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	Mode	DIALANG
1 004445 <sup>5</sup>	A2	A2	A2	A2/B1	A2	A2	A2	A2	A2	A2	A2 ↑	A1
2 004183	A2	B1	B1	B1	B1	A2	B1	B1	B1	B1	B1 ↑	A1
3 004144	A1	B1	A2	B1	A2	A2	B1	B1	B1	B1	B1 ↑	A1
4 007657	A2	A1	B1/B2	A1	A1	A2	B1	A2	A2	A2	A2 =	A2
5 004021	B2	B1	C1	B2	B2	B2	B2	B2	B2	A2	B2 ↑	A2
6 004022	B1/B2	B1	B1	B1	B2	B1	B1	B2	B2	A2	B1 ↑	A2
7 007430	B2	B1	C1	B2	A2	B2	B2	B2	B2	A2	B2 ↑	A2
8 004023	C	B2	C1/B2	B2	B2	C1	C2	C1	C1	B1	C1 ↑	B1
9 004444	A2	B1	B1	B1	B1	B1	B1	B1	C1	B1	B1 =	B1
10 008026	B2	B2	B2	B1	C1	C1	B2	C2	C2	B1	B2 ↑	B1

After rating DIALANG items, the Team members rated five Cambridge ESOL Reading tasks<sup>6</sup> as a whole. The individual items were, however, judged first, before arriving at the overall level. The tasks included: Chinese Music in an English Village (Item 1), Reading Signs (Item 2), Working in a Museum (Item 3), Natural Books (Item 4) and Advertising on Trial (Item 5). The results can be observed in Table 9.2.

After the standardisation process had taken place and a number of calibrated tasks had been rated to the CEFR levels, the Team members started to rate Matura<sup>7</sup> and The National Assessment in English for Primary School<sup>8</sup> tasks. This article will present only a few ratings of the Team members. *Matura in English Language* as The External Secondary School-

<sup>5</sup> These numbers refer to the numbers of the Dialang Reading items that can be downloaded from <http://www.lancs.ac.uk/fss/projects/grid/training/tests/reading/dialang-english-reading.pdf>

<sup>6</sup> Source: [http://www.coe.int/T/DG4/Portfolio/?L=E&M=/main\\_pages/illustrationse.html](http://www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/illustrationse.html)

<sup>7</sup> English Matura refers to The External Secondary School-leaving Examination in English that Slovenian grammar school students have to pass in order to be able to continue studying at a university.

<sup>8</sup> Primary School in Slovenian education context refers to pupils from age 6 to age 14.

leaving Examination has been assumed to be at Level B2/C1 of the CEFR levels. These claims have not been verified by any qualitative or quantitative analysis. Given this assumption, it is somewhat surprising that the items/tasks in Table 9.2 were rated much lower than the levels predicted for this examination.

*Table 9.2 Ratings for Cambridge ESOL Reading tasks by ten judges*

Test item	J 1	J 2	J 3	J 4	J 5	J 6	J 7	J 8	J 9	J 10	Mode	ESOL
1	A2	A2	B1	A2	A2	B1	A2	A2	A2	A2	A2	A2
2	A2	B1	B1	B1	B1	B1	B1	B1	B2	B2	B1	B1
3	B2	B2	B2	B2	B2	B2	B2	C1	C1	B2	B2	B2
4	C1	C1	C1	C1	C1	C1	C1	C1	C2	C1	C1	C1
5	C2	C2	C2	C2	C2	C2	C2	C1	C2	C2	C2	C2

*Table 9.3 Ratings of the 10 Matura reading items by ten judges*

Test item	J 1	J 2	J 3	J 4	J 5	J 6	J 7	J 8	J 9	J 10	Mode
1	A2	A2	A2	B1	B1	A2	B1	B1	A2	A2	A2
2	B1	A2	B1	B1	B1	A2	A2	B1	A2	B1	B1
3	B1	A2	B1	A2	B1	A2	A2	A2	A1	A2	A2
4	B2	B1	A2	B2	B1	B2	B2	A2	B2	B2	B1
5	B1	B1	A2	A2	A2	B1	B1	A2	A2	B1	A2/B1
6	B2	B1	B2	A2	B2	B1	B1	B2	B1	B1	B1
7	B2	A2	A2	A2	B2	B2	A2	A2	B1	B1	B1
8	B1	A2	B2	B1	B1	B1	B1	B2	A2	B1	B1
9	A2	B1	B2	B1	B2	B1	A2	C1	A2	B1	B1
10	A2	B1	B2	B1	B1	A2	B1	B1	A2	B1	B1

The Team members offered the following explanation for rating the tasks in Table 9.3:

- Input texts were perceived by the Team members to be authentic, relatively complex and quite long. However, the items were quite easy, for example, sometimes only “lifting” of a few words from the text was demanded for short answers, or the answer was so obvious that few or no students were likely to answer it incorrectly.
- The test methods included a short answer and a matching task. The team members noticed that especially answers in the short answer task were easy to spot and had no real distractors in the input text.
- The Team members reported that the Cambridge ESOL and the DIALANG reading items that they had judged seemed to be more demanding, had more appropriate distractors, which made the reader carefully read the text which consequently led to a better checking of the reader’s comprehension and not just understanding or even lifting a few words from the text.

The report on English Matura 2006, which is available on the website of The National Examination Centre, provides some data on how 7428 students who had taken the exam performed on the reading tasks: mean proportion correct was .69, average discrimination index +.31 and standard deviation 4.45. These measures show average students got 70% of the reading items correct.

It is to be noted that although the Team members (= judges) perceived the items to be on A2 to B1 level, the students did not find them too easy. It is therefore of vital importance that Matura Examination in English is related to the CEFR levels in the near future in a professional way so that unverified claims of the levels of the examination will not be made.

### **Designing of the Proficiency Levels**

The next step for the Team was to develop language proficiency levels for primary (lower-secondary) and *Matura*-level students. This revealed a number of questions that the group of language experts had to tackle and therefore decided to design a new set of language proficiency levels. These include factors influencing the levels, strategies characterising specific language skills etc. However, the levels have kept the main features of the CEFR levels and were mapped onto the latter (see Appendix 5). The Dutch CEF grid played an important role in the process of developing the new English language proficiency levels, especially by identifying the horizontal level of the language competence, which has been unduly ignored.

### **Piloting the proficiency levels with teachers and students**

When the proficiency levels had been finalised by the Team members, they were discussed and commented on in the study groups of English language teachers across the country and the results were included in the revision process of the levels (see Appendix 6). All descriptors for each language skill followed the same procedure, first each teacher received a chart, e.g. for reading comprehension (see Appendix 6), without any CEFR levels included as this might have influenced the teachers' decision. The results were then discussed within smaller groups and finally in the whole group. When the consensus was reached, either a plus or an x was inserted in the grid by the study group leader. The completed chart was then emailed to the Team members for consultation. The Team received about 10 charts for each language skill and more than 50% of English language teachers in Slovenia took part in this procedure. The data obtained in this way were extremely useful as they provided the Team with the information from the practising teachers on what their students can or cannot do.

### **Mapping the new proficiency levels to the CEFR levels**

In order to relate the new language proficiency levels to the CEFR the Team used a mapping procedure. This involved looking simultaneously at the CEFR scales and the designed levels and colouring descriptors that were identical (or similar) and coding those that were different. If more than 70% of the descriptors in the new designed proficiency scale were the same as in the CEFR global, analytical and/or DIALANG self-assessment scales, then the particular CEFR level was accepted. If less than 70% of the descriptors matched a particular

CEFR level, then the mapping procedure was carried out with one level lower from the previously selected CEFR level.

### **The outcomes**

In most educational settings the main reference document for language learning, teaching and assessment is the language curriculum. It is therefore a surprising fact that standard setting procedures mainly refer to language examinations and exceptionally to language curricula. I believe that language proficiency levels should be part of language curricula and a reference tool for designing language examinations and not vice versa. Such levels also suggest a way toward greater professionalism in language classroom assessment, as well as curriculum design.

The process helped to uncover many issues that would have otherwise gone unnoticed and unresolved. First, the Team members as well as many practicing teachers attending the study groups realised that designing descriptors and levels is a demanding and time-consuming task and has to be tackled more professionally. Second, understanding of the CEFR reference levels (descriptors) does not happen quickly. On the contrary, this process needs time and practice. Third, studying the CEFR (horizontal and vertical levels) is vital if one wants to design appropriate proficiency levels that would hopefully have positive a washback effect on classroom teaching. Furthermore, the Dutch CEF Grid was recognised as a very useful tool for designing levels for reading and listening as it supported the horizontal level of the CEFR and made the developers consider dimensions that would otherwise have not been observed. In addition, developing context specific proficiency levels is beneficial for the primary and secondary school teachers and above all for the students as the fairness of classroom assessment can be expected to improve. Next, relating test tasks to the CEFR makes the test designers reconsider their tasks and may lead to improvement of the test as a whole.

And finally, there is a danger that in some educational contexts, a CEFR level of an examination or a group of students at the end of a certain period of learning a language may be determined in advance or without any evidence that students claimed to be at that level can really satisfy the level's criteria (descriptors). Such procedures as described in this article may help to show how claims arrived at by mere intuition are not adequate but need to be validated.

# References

- Brumfit, C.J. and K. Johnson, ed. (1979). *The communicative approach to language teaching*. Oxford: Oxford University Press.
- Canale, M. and M. Swain (1980). "Theoretical bases of communicative approaches to second language teaching and testing". *Applied Linguistics* 1:1-47.
- Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching and Assessment*. Cambridge: Cambridge University Press.
- Grognet, A.G. and Crandall, J. (1982). *Competency based curricula in adult ESL*. ERIC/CLL New Bulletin 6: 3-4.
- Halliday, M. A. K. (1970). Language structure and language function. *New Horizons in Linguistics*, John Lyons (ed.), 140-164. Harmondsworth, England: Penguin.
- Hymes, D. (1972). "On communicative competence" in J.B. Pride and J. Holms (eds.) *Sociolinguistics*. Harmondsworth: Penguin Books.
- Van Ek, J. and J.L.M. Trim (1998). *Threshold 1990*. Cambridge: Cambridge University Press.
- Van Ek, J.L. and L.G. Alexander (1975). *The threshold level in a European unit/credit system for modern language learning by adults*. Oxford: Pergamon.
- Richards, J.C. (2001). *Curriculum Development in Language Teaching*. Cambridge: Cambridge University Press.
- Savignon, Sandra J. (1983). *Communicative competence: theory and classroom practice; texts and contexts in second language learning*. Reading, MA: Addison-Wesley.
- Schneck, E. A. (1978). A guide to identifying high school graduation competencies. Portland: Northwest Regional Educational Laboratory.
- Schneider, G. and P. Lenz (2001). *A Guide for Developers of European Language Portfolios*. Strasbourg: Council of Europe.
- Wilkins, D. A. (1976). *Notional syllabuses*. Oxford: Oxford University Press.



## Appendix 1a

Chart 1: Fill in the following chart with the appropriate information. Consider the existing English language curriculum, classroom assessment instruments, national exams, the current textbook, your teaching experience etc.

	READING	LISTENING	WRITING	SPEAKING
<b>Primary School Year 6 &amp; 9</b>	Texts: Tasks: Length: Topics: ....	Texts: Tasks: Length: Topics: ....	Narrative: Descriptive: Report: Discursive: .... Length:	Tasks: Support: Dialogue: Pair work: Monologue: ....
<b>Secondary School Year 2 &amp; 4</b>	Texts: Tasks: Length: Topics: ....	Texts: Tasks: Length: Topics: ....	Narrative: Descriptive: Report: Discursive: .... Length:	Tasks: Support: Dialogue: Pair work: Monologue: ....

## Appendix 1b

A filled-in chart by one of the Team members

	READING	LISTENING	WRITING	SPEAKING
<b>Year 6</b>	<i>Kangaroos: 110 words, Short answers Meals: 200 words, Complete the table Hobbies &amp; Games: 75 words, Answer the questions Postcards: 150 words, Match the postcards to the countries and pictures Puppy: 90 words, Form questions using key words Homes in the UK: 170 words, True/False Whales: 200 words, Short answers, T/F</i>	<i>In a pet shop: Tick the animals mentioned Wild animals: 120 words, Listen to the description and match it to the animal. Talking about the weather: tick the expressions mentioned in the text. Tick/cross in the table for can/can't Football: 70 words, Listen and put the pictures in the right order</i>	<i>Narrative: Write a dialogue about pets Look at the pictures and describe Pippi's day  Descriptive: Describe your room (60 words), Describe the weather today Write about your favourite singer, actor/actress  Report: Write about your classmates' hobbies. Fill in the chart about animals and then write about one of them  Discursive: none</i>	<i>Dialogue: Ask six classmates about their hobbies, Practise the following dialogue with your partner Practise a telephone conversation  Pair work  Monologue: Describe one of your meals Tell what the persons in the picture are doing An oral presentation about a student's pet or any wild animal</i>
<b>Primary School</b>				

## Appendix 2a

Chart 2: Fill in the following chart with the appropriate information. Consider the existing English language curriculum, classroom assessment instruments, national exams, the current textbook, your teaching experience etc. (only excerpts of the charts are shown)

### Listening Comprehension

Text type		Primary school Year 6, Age: 11	Primary school Year 9, Age: 14	Secondary school Year 2, Age: 16	Secondary school Year 4, Age: 18
Mainly Argumentative	comments formal				
	argumentation				
Mainly Descriptive	impressionistic descriptions				
	technical descriptions				
Mainly Expository	definitions				
	explications				
	outlines				
	summaries				
	interpretations				
Mainly Instructive	personal instructions				
Mainly Narrative	stories, jokes, anecdotes				
	reports				
Mainly Phatic	e.g. establishing communication, chatting, small talk				
<b>Text Source</b>					
<b>Domain</b>					
<b>Nature of content</b>					
<b>Vocabulary</b>					
<b>Grammar</b>					
<b>Text length (min/sec)</b>					
<b>Text speed</b>					
<b>Number of participants</b>					
<b>Accent/standard</b>					
<b>Clarity of articulation</b>					
<b>How often played</b>					

(Source: The Dutch CEF Grid)

## Appendix 2b

A part of Chart 2 filled in by one of the Team members

Reading Comprehension	TEXT SOURCE				Authenticity
Primary school Year 6, Age: 11					
Primary school Year 9, Age: 14					
Secondary school Year 2, Age: 16					
Secondary school Year 4, Age: 18	Advertising material Blackboard text Brochures Business letter Computer screen text Dictionaries Exercise materials	Instructional manuals Instructional material Job description Junk mail Leaflets graffiti Notices	Magazines Menus Newspapers Notices, regulations Novels OP text Personal letters Programmes Public announcements & notices Recipes	Reports Sign posting Teletext Textbooks, readers Tickets, timetables Videotext Visiting cards	Texts are authentic, no adaptation or simplification, however no specialised articles or demanding literary texts.

## Appendix 3a

Text source appropriate for a certain age group agreed on by the whole Team

<b>Text Source</b>	<b>Primary school Year 6, Age: 11</b>	<b>Primary school Year 9, Age:14</b>	<b>Secondary school Year 4, Age:18</b>
Abstracts	/	? (if adapted)	+ (no specialised language)
Advertising material	+	++	+
Blackboard text	++	+++	+
Brochures	+	++	+
Business letter	/	/	+ (only semi-business letters)
Computer screen text	+	++	+
Contracts	/	/	/
Dictionaries	+	++	+
Exercise materials	+++	+++	+
Guarantees	/	/	+ (only for objects/processes that are within students' interests and understanding)
Instructional manuals	/	/	+ (only for objects/processes that are within students' interests and understanding)
Instructional material	+	+	+ (only for objects/processes that are within students' interests and understanding)
Job description	/	/	+
Journal articles	/	/	/
Junk mail	/	+	+
Labeling and packaging	+	++	+
Leaflets, graffiti	+	+	+
Life safety notices	+	+	+
Magazines	+	++	+
Menus	+	+	+
Newspapers	/	+ (only selected texts, e.g. ads, weather forecast etc.)	+ (except for very specialised articles)
Notices, regulations	/	+	+
Novels	/	/	+ (only those that are within students' interests and understanding)
OHP text	+	+	+
Personal letters, e-mails	+	++	+
Programmes	+	+	+
Public announcements & notices	+	+	+
Recipes	+	+	+
Reference books	/	+ (only simplified or adapted texts)	+
Regulations – e.g. school, classroom rules	+	+	+
Report, memorandum	/	+ (only for objects/processes that are within students' interests and understanding)	+ (only for objects/processes that are within students' interests and understanding)
Sacred texts, sermons, hymns	/	/	/
Sign posting	+	+	+
Teletext	/	/	+
Textbooks, readers	++	++	+
Tickets, timetables	/	+	+
Videotext	+	+	+
Visiting cards	/	/	+

## Appendix 3b

Domains– appropriate for a certain age group

Domain	Primary school Year 6, Age: 11	Primary school Year 9, Age:14	Secondary school Year 4, Age:18
<b>Personal:</b> Domain in which the person concerned lives as a private individual, centres on home life with family and friends and engages in individual practices such as reading for pleasure, keeping a personal diary, pursuing a special interest or hobby, etc.	+	+	+
<b>Public:</b> Domain in which the person concerned acts as a member of the general public or of some organisation and is engaged in transactions of various kinds for a variety of purposes.	+	+	+
<b>Occupational:</b> Domain in which the person concerned is engaged in his or her job or profession.	/	+ (partly, if referring to the students' own experiences)	+ (partly, if referring to the students' own experiences)
<b>Educational:</b> Domain in which the person concerned is engaged in organised learning, especially but not necessarily within an educational institution.	+ (only if appropriate for the age group)	+ (only if appropriate for the age group)	+ (except for domains that are beyond students' own experiences)

## Appendix 3c

Nature of content – appropriate for a certain age group

Nature of content	Primary school Year 6, Age: 11	Primary school Year 9, Age:14	Secondary school Year 4, Age:18
Only concrete content	+		
Mostly concrete content		+	
Fairly abstract content			+
Mainly abstract content			

## Appendix 4

Insert A1, A2, B1, B2, C1 or C2 in the left-hand column. Do not use any documents and do not discuss the descriptors with your colleagues (excerpts only).

Level	Descriptors for Writing
	I can describe plans and arrangements.
	I can fill in forms with personal details.
	I can evaluate different ideas and solutions to a problem.
	I can describe the plot of a book or film and describe my reactions.

## Appendix 5

### The Reading Proficiency Levels for Secondary School (Year 2 and 4)

91011

Dimension	Secondary Year 2 (Age: 16)	Secondary Year 4 (Age: 18)
	Expected CEFR level: B1	Expected CEFR level: B2
<i>The student:</i>		
<b>Text source</b>	<i>reads and understands short literary (fiction and poetry) and factual texts which refer to the experience of the students and are appropriate for them<sup>9</sup></i>	<i>reads and understands longer and shorter literary (fiction and poetry) and factual texts with various topics which may not be directly related to the students' experiences<sup>10</sup></i>
<b>Topics and nature of content</b>	<i>understands texts which refers to the students' personal and everyday experiences, their personal and social environment, with concrete and semi-abstract and/or technical content.</i>	<i>understands texts which refers to the students' personal and everyday experiences, their personal and social environment, with concrete and abstract and/or technical content.</i>
<b>Organisation of the text</b>	<i>understands texts that are reasonably well structured.</i>	<i>understands texts that are complex and is able to infer the meaning from the context.</i>
<b>Reading processes</b>	<i>reads and understands details and main ideas in adapted and semi-authentic texts.</i>	<i>Reads and understands details and main ideas in authentic texts.</i>
	<i>understands facts, most arguments, facts, attitudes and significant points of the text.</i>	<i>understands arguments, facts, attitudes, and significant points of the text<sup>11</sup></i>
	<i>understands differences between main and detailed information (facts and attitudes) in descriptive and narrative texts</i>	<i>understands differences between main and detailed, explicitly and implicitly stated information different types of texts</i>
	<i>can predict and infer about the meaning of the words/phrases from the context by using different strategies</i>	<i>can quickly recognise and infer the meaning of the new words/expressions from the context</i>
<b>Support</b>	<i>can find relevant information, using a dictionary, in non-professional but reasonably complex texts to fill in information gaps; can understand the content and do the task if given time for more readings.</i>	<i>can find relevant information, using a dictionary, in professional and fairly complex texts to fill in information gaps;</i>
<b>Literary texts</b>	<i>understands shorter adapted literary text on different levels (events from the characters' perspective, theme/s, text structure, genre etc)</i>	<i>understands authentic literary text on different levels (events from the characters' perspective, theme/s, text structure, genre etc) and is able to interpret the text independently</i>
<b>Vocabulary and Grammar</b>	<i>reads and understands texts with a reasonably wide range of vocabulary and in school pre-learnt grammatical structures with a few individual idiomatic structures</i>	<i>reads and understands texts with a wide range of vocabulary and grammatical structures and with some idiomatic structures</i>

<sup>9</sup> Texts from the Internet, newspaper articles (factual, popular science and semi-professional), texts adapted for learning foreign languages, literary texts and/or their extracts (prose, drama, poetry), leaflets, instructional materials, notices, descriptions of people, places, countries etc., warnings, sign postings, formal and informal letters, posters, abstracts, recipes, advertisements, dictionary definitions, TV/radio programmes, cartoons, jokes, notes, stories and biographies.

<sup>10</sup> Texts from the Internet, newspaper articles (factual, popular science and semi-professional), texts adapted for learning foreign languages, literary texts and/or their extracts (prose, drama, poetry), leaflets, instructional materials, notices, descriptions of people, places, countries etc., film and book reviews, warnings, sign postings, formal and informal letters, posters, abstracts, recipes, advertisements, dictionary definitions, TV/radio programmes, cartoons, jokes, notes, stories and biographies.

<sup>11</sup> The students should be exposed to different discourse type texts, such as argumentative (comments, formal argumentation), descriptive, impressionistic and technical descriptions, expository (definitions, outlines, summaries, interpretations), instructive (personal instructions) and narrative (stories, jokes, anecdotes, reports). The student reads and understands these texts if they are appropriate to the students' interest, their knowledge and are not written for a very specific (too scientific, technical or cultural) audience and purpose.

## Appendix 6

Insert a plus (+) for all descriptors (objectives) that your students have achieved at the end of triad 2 or triad 3. If the descriptor is not achieved at a certain triad, insert an x.

Triad 2 Age: 11	Triad 3 Age: 14	OVERALL READING COMPREHENSION
		I can understand in detail lengthy and complex scientific texts, whether or not they relate to my own field
		I can appreciate the finer subtleties of meaning, rhetorical effect and stylistic language use in critical or satirical forms of discourse.
		I can critically appraise classical as well as contemporary literary texts in different genres.
		I can readily appreciate most narratives and modern literary texts (e.g., novels, short stories, poems, plays)
		I can make effective use of complex, technical or highly specialized texts to meet my academic or professional purposes.
		I can understand contemporary and classical literary texts of different genres (poetry, prose, drama).
		I can read texts such as literary columns or satirical glosses where much is said in an indirect and ambiguous way and which contain hidden value judgements.
		I can recognise different stylistic means (puns, metaphors, symbols, connotations, ambiguity) and appreciate and evaluate their function within the text.
		I can understand texts written in a very colloquial style and containing many idiomatic expressions or slang.
		I can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings.
		I can understand complex factual documents such as technical manuals and legal contracts
		I can recognise plays on words and appreciate texts whose real meaning is not explicit (for example irony, satire).
		I can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning.
		I can understand long complex instructions, for example for the use of a new piece of equipment, even if these are not related to my job or field of interest, provided I have enough time to reread them.
		I can understand in detail highly specialized texts in my own academic or professional field, such as research reports and abstracts
		I can go beyond the concrete plot of a narrative and grasp implicit meanings, ideas and connections.
		I can read contemporary literary texts with ease.
		I can understand in a narrative or play the motives for the characters' actions and their consequences for the development of the plot.
		I can read straightforward factual texts on subjects related to my field and interests at a satisfactory level of understanding.
		I can read straightforward factual texts on subjects related to my field of interest with a reasonable level of understanding.
		I can understand 'typical' texts that tell facts about themes I know well, e.g. short match reports, short magazine articles, factsheets, interviews with stars.
		I can follow the plot of clearly structured narratives and modern literary texts.
		I can understand the plot of a clearly structured story and recognise what the most important episodes and events are and what is significant about them.
	+	I can understand short simple messages and texts containing basic everyday vocabulary relating to areas of personal relevance or interest.
	+	I can understand short narratives about everyday things dealing with topics which are familiar to me if the text is written in simple language.
+	+	I can understand simple forms well enough to give basic personal details (e.g., name, address, date of birth).
+	+	I can pick out familiar names, words and phrases in very short simple texts.





# 10 Investigating the Relationship between the EIKEN Tests and the CEFR

Jamie Dunlea and Tomoki Matsudaira, The Society for Testing English Proficiency (STEP)


## Introduction

This report describes a standard-setting workshop that was carried out in Japan based on procedures outlined in the preliminary pilot version of the Manual for Relating Exams to the CEF. The workshop is a part of an ongoing project to investigate the relationship between the EIKEN tests in Japan and the Common European Framework of Reference for Languages: Learning, Teaching, and Assessment.

## Overview of the EIKEN Tests

Table 10.1 gives a brief overview of the EIKEN tests. There are seven levels, called grades, and each grade is administered as a separate pass-fail test. For Grades 3 through 1, the test is administered in two stages, with test takers who pass the first-stage, written test required to sit and pass a face-to-face speaking test to achieve certification at that level. The tests are produced by the Society for Testing English Proficiency (STEP). For more information on the tests, including examples of past items, readers are referred to the English version of the STEP website, [stepeiken.org](http://stepeiken.org).

Table 10.1 EIKEN tests

EIKEN Grade	LEVEL	Stage 2 (Speaking)	Recognition / Uses
Grade 1		Yes	International admissions to graduate and undergraduate programs
Grade Pre-1		Yes	
Grade 2		Yes	*MEXT benchmarks for high school graduates
Grade Pre-2		Yes	
Grade 3		Yes	*MEXT benchmark for junior high school graduates
Grade 4	Beginner	No	
Grade 5		No	

\*Ministry of Education, Culture, Sports, Science and Technology

## **Purpose and Focus of the Project**

The main aim of the project is to investigate the possibility of using the CEFR as a communication tool to talk about the EIKEN tests with language teachers and learners outside Japan. The EIKEN tests are closely integrated into the educational and social framework in Japan. In order to facilitate understanding of the EIKEN tests amongst educators outside Japan, however, it is important to describe the tests in terms that are familiar and accessible. The project will also be beneficial to Japanese learners and educators by offering them a broader perspective on what these widely used “local” measures of English language proficiency mean in terms of an international framework such as the CEFR.

The project is not designed to “link” or “align” the EIKEN tests to the CEFR. These terms may be appropriate within Europe, where the investigation of a relationship is sometimes followed by a revision process to align tests more closely with the CEFR. However, “linking” implies a stronger claim to equivalence than is appropriate for our context. The CEFR was designed and validated within Europe as a multilingual framework. The EIKEN tests have been developed and widely used for and in the EFL context of Japan. If differences between these two systems are found to exist, they may indeed be legitimate for their respective contexts, not requiring changes to either system.

## **The Standard-setting Workshop**

The workshop focused on the first-stage, written test components of the two most advanced grades, Grade Pre-1 and Grade 1. The workshop was carried out over two weekends in December 2007. The first-stage tests contain vocabulary, reading comprehension, and listening comprehension sections as well a constructed response writing task. This paper focuses only on the vocabulary, reading, and listening sections.

## **The Judges**

Thirteen judges took part. Judges were required to have at least three years experience teaching English at the university level in Japan and to have knowledge of and experience with the EIKEN tests. It was expected that few teachers would have direct experience of the CEFR, and the post-questionnaire survey proved this to be the case.

## **The Preparation Booklet**

In order to allow as much of the face-to-face time available during the workshop to be used for discussion, training, and actually rating EIKEN items, a self-study preparation booklet was provided to allow the judges to conduct the familiarization stage prior to the actual workshop. The 34-page booklet described the focus of the project and the background, purpose, and content of the CEFR. A total of eight familiarization tasks were adapted from the Manual for self-study. All relevant information to evaluate their own performance, such as the relevant scales from the CEFR, was provided.

## **The Choice of Standard-setting Methods**

Two standard-setting methods were used. The first is the variation of the yes-no method

described in the Manual, in which judges place each item into a CEFR level. The wording used for this method was: “At which CEFR level can a test taker already answer the item correctly?” The method is intuitively easy to grasp and this certainly makes training and the actual judgment process easier. For these reasons, it seems to be the method of choice for many CEFR alignment projects and is commonly referred to as the Basket method (though the Manual notes that it is in fact a variant of yes-no procedures which are very close to the original method described by Angoff).

Despite its practicality and apparent widespread use in CEFR linking projects, the Basket method remains relatively new. Variations of the Angoff method which usually involve making probability judgments remain the most widely applied methods in standard-setting (Cizek & Bunch, 2007; Cohen, Kane, & Crooks, 1999). The probability-based Angoff methods, however, have been criticized for the perceived difficulty that judges have in conceptualizing minimally competent test takers and making probability estimates (Cizek & Bunch, 2007). It was decided to use both methods to maximize their strong points. Comparing the two methods also allowed us to calculate intra-rater reliability. For the modified Angoff method for Grade 1, the judges were asked “For 100 test takers minimally competent at CEFR level C1, how many will correctly answer the item.” For Grade Pre-1, the question was “For 100 test takers minimally competent at CEFR level B2, how many will correctly answer the item.”

A deliberate decision was taken to give more weight to the modified Angoff method, because of its widespread recognition and use in standard setting. The Basket method was used first, as a “primer,” to help judges form an initial impression of items in terms of the CEFR before using the more conceptually difficult Angoff. For each grade, judges first judged items using the Basket method. After that, they then rated the same items using the Modified Angoff procedure. Finally, they were given feedback in the form of the actual proportion correct for each item when it was administered in a live test, and given the chance to change their judgments for both Basket and Angoff ratings. The EIKEN items used each came from a complete set of items administered as a live test for that grade.

## Results

Table 10.2 gives the average of ratings given by judges in each method for Grade 1 and Pre-1. For Grade 1, which is an advanced level, general English test recognized for graduate-level admissions, the point of interest was set at the cut-off between B2 and C1. Pre-1 is a lower-advanced level, general English test which is widely used for undergraduate admissions, and the point of interest was set between B1 and B2. The results show very close agreement between the cut-off points derived from the two methods. As a percentage of the total potential score for each section, the cut-off points tend to be lowest for Vocabulary, ranging from 55% to 58% and highest for Listening, ranging from 59% to 66%. For Reading the range is 58% to 61%. It is important to note that to actually pass the EIKEN tests, a score of 70% is required for both Grades 1 and Pre-1. The results allow us to make a tentative claim that test takers who have passed Grade 1 can be considered to be more than minimally competent at a level considered by the judges to be equivalent to C1.

Similarly for Pre-1, test takers who have passed the test can be considered to have demonstrated a strong performance at the B2 level, beyond minimally competent at this level. This is a very important result in terms of understanding how the content of the EIKEN tests may indeed relate to the levels described by the CEFR. However, it should be stressed that this is a tentative interpretation of preliminary results. More detailed internal validity checks on the data, along with content information from the specifications stage will need to be provided to support these claims.

*Table 10.2*

Section	Grade 1 Boundary of B2/C1			Boundary of B1/B2		
	Score for Section	Basket	Angoff	Score for Section	Basket	Angoff
Vocabulary	25	14.6	14.6	25	13.9	14.5
Reading	26	15.9	15.1	26	16.6	15.4
Listening	34	22.2	21	34	22.5	19.6

Table 10.3 shows the reliability of ratings is high across all sections. Although the Angoff method provided higher reliability for both grades for Listening, the results are mixed for the other sections. The results for intra-rater reliability for the two methods are not as high as the close relationship between cut-off scores in Table 10.2 would indicate, and we will need to investigate variation across raters in order to understand where different judges showed variation in their application of the two methods.

*Table 10.3 Reliability of ratings*

Grade	Method	Listening		Reading		Vocabulary	
		B	A	B	A	B	A
Grade 1	α	.906	.952	.874	.926	.968	.961
	Method	B	A	B	A	B	A
Grade Pre-1	α	.941*	.961	.935	.886	.940	.876

\*Based on 12 raters as R11 gave the same rating for all items

*Table 10.4 Mean intra-rater correlations for Basket/Angoff ratings*

Listening	Reading	Vocabulary	All
.789*	.722	.805	.749

\*Based on 12 raters because correlation for R12 could not be calculated

Mean inter-rater correlations are lower than we would perhaps like, but certainly not lower than many of the results reported for recent CEFR linking projects. There may be several reasons for the results, with implications for similar projects. It is of course possible that more face-to-face time for familiarization and training regarding the CEFR, and also for the standard-setting methods used, would result in higher inter-rater reliability. At the same time, we faced very severe constraints on available time. It proved difficult to get enough qualified judges able to devote more than three full days to the workshop. To have extended the workshop would have resulted in a dramatic drop in the number of judges who could attend all sessions. Our response was to provide self-study booklets. The inter-rater correlations may indicate that these were not as effective as we would have liked. Questionnaire results from judges, however, indicate that they felt the booklet and preparation tasks were indeed very useful and helped them prepare. (Readers are referred to the full PDF of the presentation made at the Research Colloquium in Athens, available on the EALTA website, for detailed results from the questionnaire).

Another possibility may be limited background knowledge of the CEFR, and this, too, has potential implications for similar projects conducted outside Europe. As stated earlier, we expected judges to have little experience of the CEFR, and this was born out in the results of the questionnaire survey. Of the 13 judges, four had not heard of the CEFR, 3 had heard of the CEFR but were not familiar with its aims or contents, and 6 were familiar with the aims of the CEFR but had not studied it in detail. Anecdotal evidence suggests that this is by no means unusual in Japan. In fact for teachers outside the university sector, where the CEFR is the focus of some attention and research, the percentage of people who have not heard of the CEFR may indeed be much higher than in this group of judges. Of course, the self-study booklet and the training with CEFR items at the workshop were designed to bring all participants to an acceptable level of knowledge regarding the CEFR. Do the results question the effectiveness of our procedures, or do they point to legitimate differences of opinion between these experienced professionals? We will need to conduct a much closer analysis of the results in order to make more informed inferences. All of the training and discussion during the workshop was video-taped, and we hope to use this resource in the future to derive a better understanding of the workshop process and the results described above.

*Table 10.5 Mean inter-rater correlations*

Grade 1	Method	Listening		Reading		Vocabulary	
		B	A	B	A	B	A
		.674	.651	.522	.514	.476	.545*
Grade Pre-1	Method	B	A	B	A	B	A
		.602	.557	.466	.412	.584	.610

\* Based on 12 raters as R12 gave the same difficulty rating for all items

## **Conclusion**

Final cut-offs have not yet be set. The project also requires addressing the speaking test components, looking at the remaining EIKEN grades, and collecting external validity evidence as well as real-world evidence on how the tests are actually being used and for what purposes. The results obtained so far have been positive, and seem to indicate that we can have confidence in the judges' decisions. Although questions remain to be answered, these questions are no less about the process of standard-setting itself than they are about the results obtained in this particular workshop. The results obtained from one standard-setting workshop, no matter how solid, would not provide incontrovertible evidence of a "link" to the CEFR. Provided that the procedures are approached in a documented, transparent manner, however, the results will provide an important source of information regarding the EIKEN tests and their relationship to the CEFR.

## References

Cizek, G., and Bunch, M. (2007). *Standard Setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage Publications.

Cohen, A., Kane, T., and Crooks, T. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education* 12(4), 343-366.

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

Council of Europe (2003). *Relating Language Examinations to the Common European Framework of References for Languages: Learning teaching, assessment. Preliminary Pilot Version*. Strasbourg: Language Policy Division.





# 11 Linking SQA's ESOL Higher to the CEFR

Rob van Krieken, Scottish Qualifications Authority

## Abstract

This paper describes attempts to link the SQA ESOL Higher Course to the CEFR scale by determining the CEFR level of exam tasks and responses. The fact that the course grade is determined by the total mark for all four skills together necessitated a practical approach. This approach consisted of describing the profile of borderline Pass and top-grade candidates. We conclude that thorough training in the CEFR is needed for all those involved in writing specifications and items as well as for those participating in linking activities.

## Purpose and context of the project

In 2006, SQA launched its new suite of ESOL qualifications. The course at Higher level intends to prepare students for employment or for study at University or College. Students vary from asylum seekers to gap-year students, and from adult established immigrants to young seasonal workers. In Scottish policy, immigration is viewed positively.<sup>1</sup>

The linking project is intended to substantiate the claim that ESOL Higher is at CEFR level C1, which is the most attractive entrance level for universities. Many similar organisations claim the same. Higher exams are located at level 6 in the SCFR (Scottish Credit and Qualifications Framework<sup>2</sup>), while Advanced Highers (equivalent to A-levels) are at level 7.

Candidates study one mandatory Unit (Everyday Communication) and one of the optional Units (Work-related contexts and Study-related contexts). Most candidates take the study-related unit. Units are assessed internally during the course, on a Pass/Fail basis, except for speaking. A course award requires passes in both unit exams. The course grade is determined by the total number of marks achieved for the external reading, listening, and writing exam papers and the internal speaking assessment.

## Design and instruments

The first linking took place in April 2007. A small panel of 4 ESOL practitioners/examiners familiarised themselves with the CEFR scales and analysed the course and unit descriptors, grade related criteria, and guidance material including examples of assessments. The panel concluded that the overall course objectives provided arguments for a C1 level, although the unit specifications were of too general a nature to give a clear indication of the required level, and the narrow range of texts and processes in the assessment exemplars were generally below C1 level.

<sup>1</sup> The 'Fresh Talent' initiative was launched in 2004.

<sup>2</sup> See [www.scqf.org.uk](http://www.scqf.org.uk) for more information.

In April 2008, a larger panel judged the 2007/08 exam and candidate evidence. The 10 judges had all been involved in the production and marking of the exam, as principal assessor, development officer, setter, or marker. Again, the panel familiarised or reacquainted themselves with the CEFR scales and discussed and judged examples before analysing the available assessments and the evidence produced by candidates.

The exam has two parts: an internal Speaking Assessment and an external exam consisting of two question papers. Skills are examined as follows:

*Table 11.1 Description of the ESOL Higher examination*

skill	questions/tasks	mins	mode	marks	mean	% correct	Alpha
speaking	2 conversations, one 2 , another 6-8 minutes	8-10	internal	25	17.27	69%	-
listening	1 mark multiple choice and sentence completion	30	external	25	17.41	70%	.77
reading	1-2 mark multiple choice, sentence completion, and short answer questions	40	external	25	15.27	61%	.69
writing	error correction 140 word letter 240 words letter (work) or essay (study)	10	external	25	15	60%	-
		30					
		50					

The cut off scores were A: 69, B: 59, C: 49, D: 44 out of 100

## Procedures and results

Because there was a lack of model fit both for the reading and for the listening assessment and the writing and speaking tasks had more marks than the maximum that can be handled by the OPLM software, the following procedure was used instead:

1. The panel judged the level of each item/task, and text, as well as the speaking and writing evidence of three grade C- and three grade A candidates;
2. CEFR levels for items and evidence were determined on the basis of majority of panel judgements, and compared with item difficulty;
3. The response patterns of the grade C and grade A candidates were described in terms of CEFR level, i.e. for each candidate the number of marks at each CEFR level was determined.

## Rater reliability

The agreement between all raters together was computed by using the multiple kappa (Gwet 2002) as well as averaging Spearman rank order correlations over all pairs of judges.

Table 11.2 gives indices of agreement between judges for each skill. Appendix 1 gives the raw judgements for reading,

*Table 11.2 Rater agreement on ESOL material by skill*

Skill 1	Multiple Kappa	Kendall's coefficient of concordance	Average Spearman rank order
Listening	0.19	0.44	0.36
Reading	0.22	0.49	0.48
Writing	0.23	0.54	0.37
Speaking	0.74	0.80	0.81

The kappa values reflect a correction for chance agreement, and tends to be lower than other indices of agreement (with only two or three relevant levels, chance agreement is high). Nevertheless, the values in Table 1 indicate that, except for speaking, agreement is very low<sup>3</sup>.

### **CEFR levels for items and evidence**

The level of each task was determined by taking the majority of judgements. Where there was no majority, the least fitting/reliable judge was removed. Table 11.3 gives the CEFR levels of reading and listening texts and items.

*Table 11.3 Number of reading and listening texts and items at various CEFR levels*

CEFR level	Listening	Reading
C1	1 text, no items	1 text, 3 items
B2	2 texts, 11 items	1 text, 16 items
B1	14 items	2 items

Table 11.4 shows how many marks A-grade candidates scored for B1 listening items etc. It is obvious that the better the grade, the higher the score is at every CEFR level. However, not even the best candidates achieve the maximum score at B1 items, while candidates with C or D grades answer some C1 reading items correctly. This reflects the wide difficulty range of the items<sup>4</sup>.

<sup>3</sup> Landis and Koch (1977) categorise 0.2 - 0.4 as fair, 0.4 - 0.6 as moderate, and 0.6 - 0.8 as substantial.

<sup>4</sup> Difficulty of B2 reading items varied from .05 to .94, for listening difficulty varied from .51 to .95 for B1 items, and from .10 to .94 for B2 items (on average, these B2 items were more difficult). This may reflect less than perfect item construction and panel judgements.

*Table 11.4 Average score for CEFR level items obtained by differently graded candidates*

Grade	Candidates	Listening		Reading		
		B1 items max=14	B2 items max=11	B1 items max=3	B2 items max=19	C1 items max=3
A	41	12.73	7.54	2.71	14.02	1.61
B	19	11.26	5.68	2.42	10.11	1.32
C	16	10.25	4.63	2.25	9.50	1.13
D	7	8.43	5.00	1.86	7.00	0.86
None	4	6.50	2.75	2.25	4.75	0.75

Table 11.5 shows how the panel judged the level of the speaking and writing evidence produced by borderline grade C and grade A candidates. Note that the tasks did not set a ceiling for the level.

*Table 11.5 CEFR levels of evidence produced by 6 borderline candidates*

Candidates	Speaking	Writing (2 pieces)
A-grade	Three C1	One C1, five B2
C-grade	One C1, one B2	Two B2, three B1-

### **The total response patterns of the grade C and grade A candidates in terms of CEFR level**

Table 11.6 finally gives the marks for items/evidence at each CEFR level of three C-grade and three A-grade candidates.

*Table 11.6 Profiles of C- and A-grade candidates in marks per CEFR level*

	Cand A1	Cand A2	Cand A3	Cand C1	Cand C2	Cand C3
<b>C1</b>	22	26	21	12	12	1
<b>B2</b>	39	29	33	21	19	37
<b>B1</b>	17	13	12	18	18	8

All candidates were able to demonstrate C1 level in speaking, but for lack of C1 items A-grade candidates could not demonstrate this level in reading and listening.

### **Summary and discussion**

The results show that the panel should have been trained better in the use of CEFR levels. In addition, the amount of speaking and writing evidence analysed was quite small. Even so, it is clear that reading and listening tasks were generally below the intended C1 level. This is due to a lack of detail in the specifications and a focus on retrieval of explicit information in a narrow range of texts. It was also noticed that items within the same CEFR

level vary widely in difficulty.

It proved to be difficult to collect independent judgements from a panel which was deeply involved and interested in reaching agreement on necessary improvements.

# References

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

Council of Europe (2003). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment: Manual, Preliminary Pilot Version*. Strasbourg: Council of Europe.

Kilem Gwet (2002). *Computing Inter-Rater Reliability With the SAS System*, Series: Statistical Methods For Inter-Rater Reliability Assessment, No. 3, October 2002.

Landis, J.R. and Koch G.G. (1977). "The measurement of observer agreement for categorical data," *Biometrics*, 33, 159-174.

McIlwraith, Hamish (2007). *Benchmarking the SQA Higher ESOL to the Common European Framework*. Internal SQA report.

## Appendix 1: raw judgements on reading

	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	total "C1"	total "B2"	total "B1"
text													
1	C1	B2		B2	B2	B2	B2	B2	B2	B2	1	8	0
2	B2	C1		C1	C1	C1	C1	C1	C1	C1	8	1	0
item													
1		B1	B1	B1	B1	B2	C1	B1	B1	B1	1	1	7
2		B2	B1	B1	B1	B1	C1	B1	B1	B1	1	1	7
3		B2	B2	B2	B1	B2	C1	B2	B1	B2	1	6	2
4		B2	B1	B2	B1	B2	C1	B2	B2	B2	1	6	2
5		B2	B1	B1	B2	B1	C1	B1	B2	B2	1	4	4
6		B2	B1	B1	B2	B1	C1	B1	B2	B2	1	4	4
7		B1	B1	B2	B2	B1	C1	B1	B2	B2	1	4	4
8		B2	B2	B2	B1	B1	B2	B2	B1		0	5	3
9	B2	B2	B1	B1	B1	B1	B2	B2	B1	B1	0	4	6
10	B2	B2	B2	B2	B1	B1	B2	B2	B1	B2	0	7	3
11	C1	B2	B1	B2	B2	B1	C1	B2	B2	B2	2	6	2
12	C1	B2	B2	B2	B2	C1	C1	C1	C1	C1	6	4	0
13	B2	B2	B2	B2	C1	C1	B2	B2	B2	B2	2	8	0
14	B2	B2	B2	B2	B2	C1	B2	B2	B2	B2	1	9	0
15	B2	B2	B2	B2	B2	B2	B2	B2	B2	B2	0	10	0
16	C1	C1	C1	C1	C1	B2	c2	C1	C1	C1	8	1	0
17	B2	B2	B2	B2	B2	C1	B2	B2	B2	B2	1	9	0
18	B2	B2	B2	B2	B2	C1	C1	B2	B2	B2	2	8	0
19	C1	B2	C1	B2	B2	C1	C1	C1	C1	C1	7	3	0
20	B2	C1	B2	B2	B2	C1	B2	B2	B2	B2	2	8	0
21	B2	C1	B2	B2	B2	C1	B2	B2	B2	B2	2	8	0





# 12 Bilkent University School of English Language COPE CEFR Linking Project

Carole Thomas and Elif Kantarcioglu, Bilkent University, Ankara

## Purpose and context of the project

### Background

Bilkent University is an English medium university in Ankara, Turkey. Students enrolled in the university but who do not meet the required proficiency level of English, attend Bilkent University School of English Language (BUSEL). BUSEL produces and administers Bilkent's proficiency exam, the Certificate of Proficiency in English Exam (COPE), which is taken each year by approximately 3000 students. It is a high-stakes exam which determines whether or not students have the language proficiency required for academic study in English.

### Goals

The COPE Linking Project was initiated in July 2006, with the overall aim of aligning the COPE to the CEFR at B2 level. It is being carried out to strengthen the validity claim of the exam and is in line with one of the principal aims of the CEFR, "to facilitate the mutual recognition of qualifications gained in different learning contexts and aid European mobility" (Council of Europe, 2001:1).

### Relevance to National Policy

The linking project was also undertaken to reflect political initiatives in Turkey, namely participation in negotiations to join the European Union and in support of the 1999 Bologna Declaration, which aims to facilitate mobility of students in the context of higher education. The country aligns itself strongly with European educational norms and practice and support for the CEFR is spreading. Several conferences with a CEFR focus have been held and the ELP has been introduced into a number of Turkish schools.

## Project design

### Scope

The COPE linking project follows the methodology described in the Manual for Relating Language Examinations to the CEFR (Council of Europe, 2003) with its four interrelated stages – Familiarisation, Specification, Standardisation and Empirical Validation.

### The COPE examination

The COPE, which measures general and academic English, consists of 4 papers; reading, writing, listening, and language and is designed, trialled, implemented and analysed in line with good test practice. The Test Specifications were revised in 2006 using Weir's socio-

cognitive framework for validating tests (2005) and standards are maintained using Item Response Theory for anchoring, post test analysis and item banking.<sup>1</sup>

### **Standard setting methods**

The Examinee-Paper Selection Method was used for the standard setting of the writing paper because it is considered to be the most suitable for polytomously scored performance tasks and also because standard setting decisions are based on actual examinee papers (Hambleton, Jaeger, Plake & Mills, 2000). Both the original Angoff, also known as the Footnote method, and the Yes/No method, an Angoff based method modified by Impara and Plake (Cizek and Bunch, 2007), were employed for the reading and listening paper standard setting. The two Angoff based methods were chosen as they are not only the most widely used and most thoroughly researched standard-setting methods but are also “well suited for tests comprising multiple choice format items” (Cizek and Bunch, 2007: 82). The use of two distinct methods allowed for potentially different cut scores and comparison of these, which then contributed to the reliability of the established cut-score.

### **Selection and training of judges**

In order to bring in different perspectives, the team of judges comprises a group of 15 people, representing the different parties in the school. It was also decided that in order to avoid bias and to have an outsider perspective on the exam, external experts would be called in at various stages of the project to act as judges.

The Manual stipulates that the judges should have an in-depth knowledge of the CEFR and suggests a range of familiarisation activities. The activities suggested in the Manual (Council of Europe, 2003: 25), formed part of the familiarisation process. These activities were carried out as suggested but were felt to be insufficient. In order to promote a deeper understanding of the CEFR, a number of additional activities were added, ranging from article discussion to an introduction to IRT and its usage in calibrating the CEFR levels. The degree of the judges’ internalisation of the CEFR was monitored until it became clear that they were ready to progress to the standardisation stage. Before the actual standard setting took place, the judges were further trained by using the CEFR calibrated samples provided by the Council of Europe.

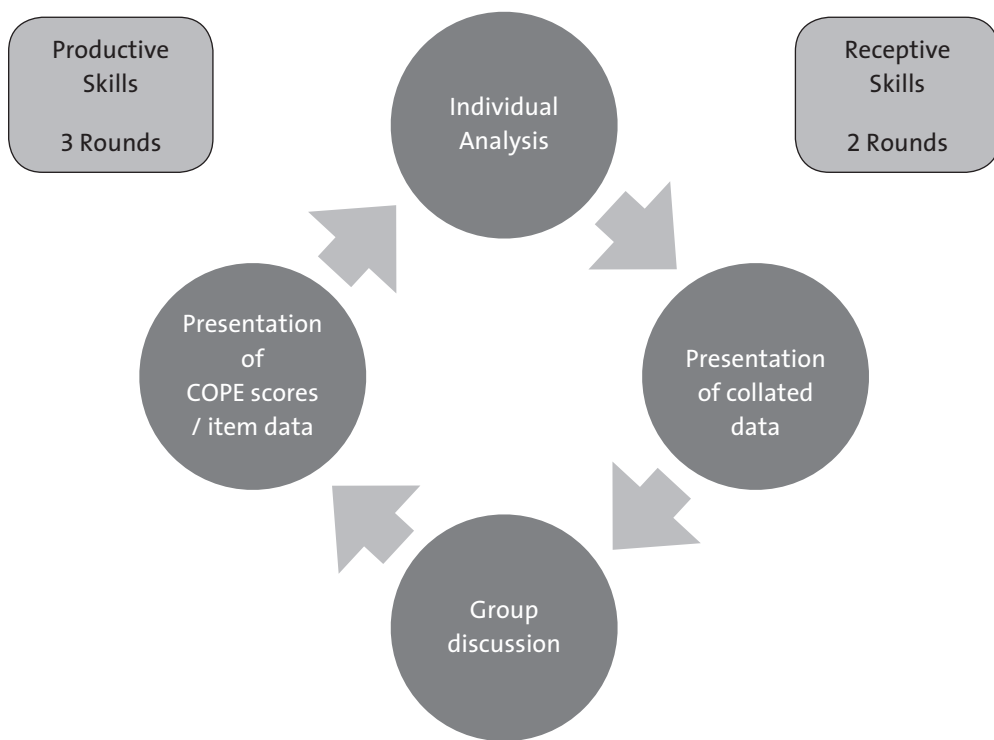
## **Standard Setting procedures and results**

### **General Methodology**

Although different standard setting methods were used for productive and receptive skills, the actual procedures followed were similar. As well as the cyclical approach followed in Figure 12.1, the judges were asked to indicate their confidence level in their judgments for each of the sample papers/items at the end of each round (Hambleton and Plake, 1995).

---

<sup>1</sup> [www.bilkent.edu.tr/BUSEL](http://www.bilkent.edu.tr/BUSEL) The full project report will be available by June 2009



*Figure 12.1 Procedures followed in the standard setting sessions*

### **Results**

The consistency of the judges was demonstrated using Cronbach alpha and Pearson correlation. The agreement among the judges was demonstrated using intraclass correlation coefficient (ICC), which shows to what extent the average rater agreed with all others. As Kaftandjieva (Council of Europe, 2004:23) highlights in the Reference Supplement to the Manual, correlational analyses are not appropriate for standard setting purposes as “it is possible to have a perfect correlation of  $\pm 1.00$  between two judges with zero-agreement between them about the levels to which descriptors, items, examinees or their performances belong.” Therefore, consistency and agreement among judges are also reported using FACETS, the program implementing the many-facet Rasch measurement model (Linacre, 1989). This model is helpful in looking at a score which is based on a number of facets. In this case, the facets involved are the samples, judges and the CEFR scales. For more details, please refer to the full report.

The data in Tables 12.1, 12.2 and 12.3 show that the standard setting sessions were successful. Although the correlation among the judges for the reading paper in particular was low, the judges expressed high confidence in their judgments on a scale of 1 to 4.

Table 12.1 – Agreement and consistency among judges

	N	Alpha	ICC	Pearson correlation	Average Confidence Level
Writing	11	.9963	.9704	.946	3.40
Reading – Yes/No	10	.7901	.7920	.313	3,44
Reading – Angoff	10	.9223	.9217	.286	3,22
Listening – Yes/No	11	.9347	.9347	.566	3.41
Listening – Angoff	11	.8946	.8946	.459	3.33

Table 12.2 – Writing

Writing	N	Mean	Median	Mode	SD	Range	Min	Max
1	11	5	5	5	0	0	5	5
2	11	7	7	7	0	0	7	7
3	11	7.35	7	7	0.49	1	7	8
4	11	8.21	8	8	0.42	1	8	9
5	11	7	7	7	0	0	7	7

Table 12.3 – Reading and Listening

	N	Mean	Median	Mode	SD	Range	Min	Max
Reading - Yes/No	10	20,5	20,5	20	2,91	7	17	24
Reading – Angoff	10	19,09	19,24	19,40	2,08	7,7	14,3	22
Listening - Yes/No	11	15.82	16	16	2,18	6	13	19
Listening – Angoff	11	14,65	14,79	-	1,72	5,79	11,61	17,4

## Summary and discussion

### Validity, reliability and generalisability of the results

The first standard setting session was carried out with the COPE writing paper, however, due to the limited number of papers analysed, no CEFR alignment claim can be made at this stage. Standard setting for the reading paper was initially problematic due to the use of several standard setting methods and working with the “least able B2” candidate profile. Three further sessions were held until the judges were confident in the use of the standard setting methods employed. The listening cut-score was the last to be set and by this time the group was very experienced and no problems were encountered. The reliability of the cut-scores for these papers will be further confirmed when the live exam is run in June 2008. Work is currently underway with a group of teachers to develop their skills in evaluating their students in terms of the CEFR and an analysis of this data will be correlated with the students’ actual scores in the COPE exam.

**Issues raised and recommendations**

It proved to be unrealistic to carry out standard setting in accordance with the schedule suggested in the Manual and the length of time needed for discussion must be reconsidered. The limited choice of samples available for standardization is also an issue, as is the fact that the academic context is under-represented. There is a need for a broader range of samples and it would be useful if IELTS or TOEFL would make available sample writing papers with their CEFR levels.

**Plans for further work**

The project will be completed by the end of 2008 once the principal aim of linking COPE to the CEFR has been achieved. However, work with the CEFR will continue through the familiarization work which was started with the project members and will eventually spread throughout the school and become a part of the institutional culture. For BUSEL, the project has been invaluable; both in terms of familiarization with the CEFR and with the work that has been undertaken to relate the B2 level to the academic context.

# References

Cizek, G.J. and Bunch, M.B. (2007). *Standard Setting*. Thousand Oaks, CA: Sage.

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

Council of Europe (2003). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment: Preliminary Pilot Manual*. Strasburg: Council of Europe, Language Policy Division.

Hambleton, R.K., Jaeger, R.M., Plake, B.S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, 24, 355-366.

Hambleton, R. K., & Plake, B. S. (1995). Using an Extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8(1), 41-55.

Linacre, J.M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Weir, C.J. (2005). *Language Testing and Validation: an evidence-based approach*. Oxford: Palgrave.

# 13 Standard Setting for Listening, Grammar, Vocabulary and Reading Sections of the Advanced Level Certificate in English (ALCE)

N. Downey and C. Kollias, Hellenic American Union & Hellenic American University

## Abstract

This paper describes the standard setting procedures conducted by the Hellenic American University and Hellenic American Union as part of the process to link the Advanced Level Certificate in English (ALCE) to the Common European Framework of Reference for Languages (CEFR). It begins with a brief overview of the ALCE and of the Linking Project, before looking in detail at the Standardization Phase of the Project. The procedures followed are described, including the method used for standard setting, the selection and training of the judges, the data obtained from the judges and the analysis of this data.

## Purpose and Context of the Linking Project

The Advanced Level Certificate in English (ALCE) is a high-stakes examination designed for candidates who require certification of their competency in English as a foreign language at an advanced proficiency level. The candidates are tested on their communicative competency in all four language skills: reading, writing, listening and speaking, as well as in their vocabulary and grammar resources. The examination is administered in Greece, Turkey and the Balkans.

The Linking Project was begun in order to assess the level of the ALCE examination in terms of the levels defined by the CEFR. The ALCE examination was designed to bridge the gap between examinations claimed respectively to be at B2 and C2 level and had been extensively revised in the light of the CEFR scales and descriptors. It was therefore necessary to investigate whether the revised ALCE examination could be linked to the C1 level.

## Design of the Linking Project and the Instruments Used

The Project involved familiarization with the CEFR and the descriptors of the levels, followed by a detailed analysis of all parts of the ALCE examination in order to determine its correlation with the levels of the CEFR. This was then followed by the Standardization Phase.

The Linking Project was carried out by a committee of ten members, comprising eight item writers and two Coordinators working for the Hellenic American University on the ALCE

examination. All Project members but one were holders of an MA in TEFL and were teachers of EFL and teacher trainers, in addition to being testing specialists. Seven of the members had a native speaker background and three were non-native speakers. Throughout the linking process, the committee was advised by consultants from Cito, Netherlands.

The Project was divided into four phases, as proposed in the *Preliminary Pilot Version of the Manual for Relating Language Examinations to the CEF* (the Manual) (Council of Europe: 2003): Phase 1: Familiarization; Phase 2: Specification; Phase 3: Standardization; and Phase 4: Empirical validation.

In the first phase, an in-depth familiarization process with the content and levels of the CEFR was carried out before proceeding with further phases. The specification phase involved the project members mapping the coverage of the ALCE examination in relation to the categories and levels of the CEFR. First, a full Manual for the ALCE examination was produced, which described the format, content and rationale of all sections of the examination in detail ([www.hau.gr/resources/pdf/alce\\_manual.pdf](http://www.hau.gr/resources/pdf/alce_manual.pdf)). Using information gained during this process, a content analysis of the ALCE examination was then carried out in order to complete Forms 1 – 23 of the Preliminary Pilot Manual.

The ALCE examination was felt to be uniformly aimed at the C1 level. This is consistent with its design and rationale as a high stakes examination at advanced proficiency level.

### **Phase 3 of the Linking Project: Standardization**

The procedures for the standard setting of the ALCE examination follow those set out in the Common European Framework of Reference for Languages, the Preliminary Pilot Version of the Manual on Relating Language Examinations to the Common European Framework of Reference for Languages, and the Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages.

The procedures used were as follows:

- the selection of the most appropriate and effective method of standard setting for each Section of the ALCE examination;
- the selection of a large number of judges, based on their qualifications and experience of language teaching and testing;
- the training of the judges on the appropriate criteria, as defined in the *CEFR* and the *Manual for the Advanced Level Certificate in English* ;
- the establishment by the judges of cut-off scores for each Section
- data analysis of the judges' cut-off scores.

Since the Listening Section and Grammar, Vocabulary, Reading Section both comprise multiple-choice items and are machine scored, the standard setting for these two Sections followed precisely the same procedures.



The standard setting was carried out on the items used in the January 2006 ALCE examination, and occurred in successive meetings during January and February 2006.

### **Selection of the most appropriate and effective method of standard setting**

Standard setting for the Listening Section and Grammar, Vocabulary, Reading Section of the ALCE examination was carried out using the modified Angoff method, defined as “judgements of the proportion of correct responses for minimally qualified examinees on each item of the test” (Reckase 2000: 3). This method was chosen due to its appropriateness for a multiple-choice format and its efficiency of use.

In carrying out this method, carefully selected judges first took part in a thorough familiarization and training session before assessing each Section. They were then asked to examine each item and assess the percentage of candidates minimally acceptable at the level that would be likely to choose the correct answer choice for this item. The judges’ decisions were collated and a cut-off score for each Section set.

### **Selection of judges**

Brandon (2004: 68) concluded that the ideal number of judges taking part in a modified Angoff standard setting procedure should be between 15 and 20. However, for greater precision in setting the cut-off score estimates, 20 judges were used for the standard setting of the Listening and the Grammar, Vocabulary, Reading Sections.

The judges used were selected for their qualifications and experience of teaching English as a Foreign Language and of testing in this field. Ten of the judges had prior experience of standard setting procedures, of which nine also had experience of item writing for high-stakes tests. All the judges were practicing teachers at the time of the standard setting and all had experience as Oral Examiners for a variety of tests at a variety of levels.

Each of the judges was asked to complete a Background Information Form. The form records a summary of their educational and teaching experience, as well as any experience they may have had with the ALCE examination, for inclusion in the documentation of the standard setting. In addition, a Curriculum Vitae of each judge was also kept on file.

### **Training of the Judges**

In order to familiarize the judges with the CEFR and the CEFR scales, first, they were introduced to the rationale and context of the CEFR. The next part of the judges’ training involved familiarization with the ‘can do’ statements of the CEFR.

Before each Section of the ALCE examination, the judges were required to examine the ‘can do’ statements for that Section, both the overall descriptors and those more specific, in order to rank them according to the six scales of the CEFR. The participants carried out this task individually and then compared their rankings in pairs and groups. A detailed discussion of the rankings followed, with participants justifying their decisions and receiving feedback on the order as defined in the CEFR. All present were able to reach a

consensus on the ranking without difficulty and developed their awareness of what defines each level as described by the CEFR. Participants were then directed to focus on the descriptors for the C1 level – the level which the ALCE examination is aimed at. This level was discussed in detail and contrasted with the C2 level above and the B2 level below in order to clarify exactly what C1 means in terms of the CEFR. For the grammar and vocabulary sections, the DIALANG scales were also used.

Having established the criteria for the level, the judges were then trained in the criteria for making a decision on actual test items according to the modified Angoff Method. They were asked to examine each of the test items in terms of the percentage of candidates minimally acceptable at the level that would be likely to choose the correct answer choice.

### **Establishment of cut-off scores**

Familiarization was carried out for Listening, Grammar, Vocabulary and Reading in turn. After each familiarization activity, the judges examined the items relating to that particular language area. For the Listening Section, a recording was played with a longer pause allowed after each item for the judges to record their decision.

Forms were given to the judges to record their decisions for each of the language areas and the same procedures were followed for each. They first recorded their individual decision and were then given the key to each item. The opportunity to discuss their findings in pairs and groups then followed, after which they once again recorded their decision, based on the discussion. The judges were then given empirical statistics on how candidates performed on the items, followed by a plenary discussion, and then they recorded their final decision. In this way, each judge recorded three impression marks for each of the items.

### **Data analysis of the cut-off scores**

The judges' decisions were collated and merged to give the overall cut-off scores for each section. The figures for items of each Section were entered into an Excel database. The median, average and standard deviation for each rater were calculated, as well as the minimum and maximum score given. In this way, intra-rater and inter-rater correlations could be examined.

The cut-score for the ALCE was calculated on the final empirical round data. The formula for the Standard Error of the cut-score (SE<sub>c</sub>) was calculated as a validity check (Council of Europe 2004: 21- 22). The results were the following:

Table 13.1 Validity check of Standard Setting cut-off scores

	Listening Section	Grammar, Vocabulary and Reading Section
Standard Deviation of cut-off score (SDc)	3.99	3.73
Standard Error in the test (SEM)	2.88	4.60
Standard Error of cut-off score (SEc)	0.89	0.83
Validity check	0.31	0.18

An average rating for each item was produced from the figures given and the average of these ratings was calculated to give the cut-off score.

The internal validity check revealed that the SEc for the Listening section and the Grammar, Vocabulary and Reading section were .31 and .18 of the SEM respectively. Cohen, Kane, & Cooks (1999: 364) claim that SEc should be at least less than .5 of the SEM to ensure minimum impact on the misclassification rates. Thus, the SEc of both sections “can be considered as relatively small and acceptable” (Council of Europe 2004: 22).

## Summary and Discussion

The cut-off score obtained from the standard setting for both Sections was lower than that expected and implies that some items in the ALCE examination may be pitched at a slightly higher level than intended.

One difficulty encountered during the standard setting process stemmed from the fact that the CEFR does not provide scales for tests that include discrete grammar and vocabulary items. To compensate, the judges were given the DIALANG scales to use for their evaluation of these items. However, the DIALANG Grammar scales mostly focus on form, whereas in the ALCE grammar section the primary focus is on meaning.

A suggested modification to any future standard setting carried out on the ALCE is to hold parallel session running with at least 10 judges in each group and then statistically compare group cut-off scores. Furthermore, it may be preferable to use a different variation of the Angoff Method, in particular the Yes/No method or Borderline method proposed by Impara and Plake (cited in Cizek & Bunch, 2007: 88-89) as this method does not require judges to enter a probability score, but to state whether the “borderline” candidate would get the item correct. The continued use of three rounds would allow the judges to get feedback on their estimates before a final cut-off score is calculated.

## References

Brandon, P. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17(1), 59-88.

Cohen, A.S., Kane, M.T. & Crooks, T.J. (1999) A generalized examinee-centered method for setting standards on achievement test. *Applied Measurement in Education*, 12(4), 343 – 366.

Cizek, G, & Bunch, M. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance on Tests*. London: SAGE

Council of Europe (2003). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF). Manual. Preliminary Pilot Version* <[www.coe.int/T/DG4/Portfolio/documents/Manual%20for%20relating%20Language%20Examinations%20ot%20the%20CEF.pdf](http://www.coe.int/T/DG4/Portfolio/documents/Manual%20for%20relating%20Language%20Examinations%20ot%20the%20CEF.pdf)> accessed 20/07/08

Council of Europe (2004). *Reference Supplement to the Preliminary Pilot version for the Manual for Relating Language Examinations to the Common European Framework of Reference Language for Languages: Learning, Teaching, Assessment (CEF)* <<http://www.coe.int/T/DG4/Portfolio/documents/CEF%20reference%20supplement%20version%203.pdf>> accessed 20/07/08.

Reckase, M. D. (2000). *The ACT/NAGB Standard Setting Process: How “Modified” Does It Have To Be before It Is No Longer a Modified-Angoff Process?* Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000). (ERIC Document Reproduction Service No. ED442825)

## 14 Jack of More Trades? Could standard setting serve several functions?

Eli Moe, University of Bergen, Norway

Although everyone agrees that standard setting is a must when linking language tests to the Common European Framework of Reference (CEFR) (Council of Europe, 2001), we hear complaints about the fact that standard setting is expensive both in respect to both time and money. In addition it is a challenging and exercise for judges, not only because the CEFR gives little guidance on what characterises items mirroring specific levels, but also because time seldom seems to increase individual judges' chance of success in assigning items to CEFR levels.

Several researchers have stressed the need for strengthening the ties between the CEFR on the one hand side and learner language and theories of second language learning on the other. A legitimate question is therefore whether unveiling features characterising items mirroring specific CEFR levels in connection with standard setting, could serve several purposes. Such knowledge could most likely help judges to target the CEFR levels items mirror more easily, and thus add to the validity of the standard setting procedures as well as to the result. At the same time it would contribute to an empirical and/or theoretical basis for the CEFR levels. The main aim of these pages is to encourage reflection on and attention to standard setting to the CEFR – the procedures, outcomes and future development.

### **The CEFR and standard setting to the CEFR**

The CEFR describes language competence at six levels: A1 – A2 – B1 – B2 – C1 – C2. While A1 refers to very basic language competence, C2 mirrors advanced academic competence. Level descriptors characterise what learners can do at each CEFR level. In a language learning context these may function as aims, something the learner works towards. In many European countries CEFR descriptors function as central aims in the language curricula. In Norway for instance, the foreign language curriculum in primary and secondary school is based on the CEFR (Utdanningsdirektoratet 2006), as well as the curriculum which is the basis of free language courses offered to adult immigrants (Vox 2005). The CEFR seems to function well in a pedagogical context, while some say it is more problematic in connection with testing and second language research (Alderson, 2007; Hulstijn, 2007; Weir, 2005). The level descriptors have come about as a result of a joint collaboration of language researchers, pedagogues, testers and psychometricians. The descriptors thus express a consensus on what characterises learner language and how it develops. It is, however, important to realise that the descriptors are only indirectly based

on learner language. This means that the CEFR is not based on actual second language data, nor on any specific theory about second language learning. During the last few years several researchers have addressed this problem and expressed a need for more research in order to strengthen the link between the CEFR and learner language on the one hand, and theories of language learning on the other.

*“The term “standard setting” in the field of educational measurement refers to a decision making process aiming to classify the results of examinations in a limited number of successful levels of achievement (Kaftandjieva, 2004: 2).”*

Standard setting procedures are applied in connection with criterion referenced tests, i.e. when the test results point to certain levels of competence. There are more than 50 different standard setting methods (Kaftandjieva, 2004:12), and according to Jaeger (1989:493) most of these belong to one of two main groups: examinee centred methods or test centred methods. Applying an examinee centred method teachers assign students' competence to competence levels. Teacher assignments together with the students' test data are the basis for establishing cut-off scores. In contrast the starting point for a test centred method is judges' assignment of items to competence levels. The cut-off scores are based on the level assignment of items as well as the empirical difficulty estimate of the items.

The basis for standard setting is on the one hand a valid and reliable test. On the other hand standard setting depends on reliable judges who are acquainted both with the competence levels the test is supposed to measure, and also how to recognise the competence level needed in order to answer actual items correctly. A number of judges will then normally assign test items or the competence of actual students taking the test to competence levels. The description of the competence levels and the preparation of the judges affect the standard setting. The clearer the description of the judges' task is, the more focussed the training of the judges can be. At the same time the probability for setting correct cut-off scores increases. Researchers agree that a positive relationship between judges' assignment of items to competence levels and observed item difficulty supports the validity of the cut-off scores established. The more positive this relationship is, the more valid the cut-off scores are.

### **The task of the judges**

An increasing number of language tests in Europe (and beyond) claim to be linked to the CEFR. Some of these tests provide evidence for linkage by reporting on standard setting procedures and results. Cut-off scores between CEFR levels are established by applying one or more standard setting methods and thus combining empirical data with human judgement. A valid and reliable test, CEFR familiarisation and training of judges represent a common basis for making decisions about cut scores between levels.

Whether choosing to apply a test centred or examinee centred method, qualified judges play a crucial part in the standard setting process. The judges' task is:

- to be well acquainted with the mental construct called the CEFR;
- to interpret levels of competence within this mental construct;
- to recognise the language competence the CEFR levels mirror in items and students;
- to recognise the language competence required to answer different items correctly and match this with CEFR level descriptions;
- to match language learner competence with the competence inherent in descriptions of the CEFR levels.

When several judges assign individual items to CEFR levels, they do this by comparing items to CEFR descriptors. To some extent this ensures the connections between items and tests on the one hand, and CEFR level descriptions on the other. But there is also a certain vagueness about such an action. Judgements are predictions about something or someone. The CEFR is a mental concept of levels of language proficiency. Assigning items to CEFR levels therefore involves making predictions in relation to a mental concept – the nature of which is not very concrete. Applying an examinee centred method on the other hand means that teachers are making predictions about students' language proficiency; i.e. teachers assess actual students who they know. This task may seem more concrete than the former. Studies involving judges or raters often show that they disagree or vary in their judgements. This is the main argument against using an examinee centred method as the only approach in standard setting studies. So both test centred methods and examinee centred methods have their strengths and drawbacks. While we can rest our final judgement on the opinions of several judges, applying a test centred method, the judges' task is more vague and less concrete. In connection with a examinee centred method we rely on one teacher's judgement about each individual student. But the teacher knows his student quite well, which makes his task more concrete and less vague.

The responsibility placed upon judges is not trivial. Cut-off scores between CEFR levels are derived more or less directly from the judges' assessment of either test items or test takers. These cut-off scores have a direct impact on real test takers. If the judges don't have a clear notion of the task they have to perform, the basis for making judgements is somewhat problematic, as are also the final cut-off scores.

### **Challenges encountered with setting CEFR standards: The Bergen experience**

At the University of Bergen, Norway, language testers are trying to work systematically in order to document that tests in Norwegian as a second language and national tests in English for Norwegian school children are linked to the CEFR. On the one hand side this has been something we have felt obliged to do as the call for standard setting has spread through out Europe. On the other hand side it has been, and still is, a very interesting journey since different tests and contexts tends to produce different problems and challenges.

Applying a test centred standard setting method involves judges who assign test items to CEFR levels. Prior to this, they have to be trained for the job they are going to do. CEFR level descriptions and typical items already assigned to levels serve as a help in this context.

When some judges turn out to be less successful than others one could ask whether the quality of the training or the individual judge is to be blamed.

Experience with linking language tests to the CEFR tells us that:

- judges find it challenging to relate items to CEFR levels;
- standard setting studies often report modest correlations between the p-values of items and judges' CEFR level assignment of items;
- some judges are better than others – in fact the same judges tend to be successful at predicting item difficulty every time standard setting procedures are repeated;
- abandoning less successful judges and keeping the most successful ones is seldom a realistic option, since new judges will have to be recruited and trained before knowing how successful they will become.

Since the CEFR describes language proficiency, and not what it takes to measure this proficiency, standard setting judges have to infer how the CEFR descriptors relate to students and items.

Standard setting judges often say that they think it is challenging to match CEFR levels and items. They say it is difficult to recognise what competence level a learner must have in order to answer items correctly. This kind of feedback is not unexpected, because the CEFR does not address this question. While the CEFR describes what learners can do at different levels, it does not say anything about what characterises items/questions mirroring different CEFR levels. A listening or reading text can be considered suitable for a level, but the items developed for such a text may not necessarily target this level. Some items may be very easy or very difficult, and target the levels below or beyond the targeted level. During standard setting items sometimes are assigned to different levels than the targeted. Standard setting judges (and the experts guiding the judges) are therefore faced with a challenge not covered by the CEFR.

## **A potential way of setting standards and strengthening the CEFR argument**

Is it possible to say anything at all about what characterises items mirroring competence at specific CEFR levels? If B1 or B2 competence exist, is it then possible to describe how this competence can be measured in a specific context? Or is it possible to find out what characterises for instance B1 items in a concrete setting? If the standard setting procedures could help unveil parts of the relationship between CEFR level proficiency and what characterises items measuring this proficiency, this would probably both strengthen the link between a test/items and the CEFR, make standard setting judges more successful in predicting item difficulty, ensure the validity of the cut-off scores established, as well as support item writers in targeting specific CEFR levels.

To describe in detail “what a B1 item is”, seems difficult, because there is an almost infinite number of possible text/question combinations. A pragmatic approach would be to start in a small corner and describe what characterises B1 items in a small limited context. Later one could add to the initial description by describing “typical” items below and above this CEFR level.



A standard setting study conducted at the University of Bergen in 2008 aims to find a core of A2 and B1 reading items to be used in further piloting and testing for two different tests in Norwegian as a second language for adult immigrants, *Norskprøve 2* (Test in Norwegian 2 (A2)) and *Norskprøve 3* (Test in Norwegian 3 (B1)) (Moe, 2008). Both a test centred method, the Kaftandjjeva & Takala Compound Cumulative Method (Kaftandjjeva & Takala, 2002), and two examinee centred methods, the Contrasting group method and the Borderline method (Cizek & Bunch, 2007), are used in this study. In the process we discovered that for three of the booklets it was impossible to establish cut-off scores between A1 and A2 and B1 and B2 applying the Contrasting group method since no, – or very few students, were assigned to A1 and B2. We therefore ended up basing our final cut-off scores using the Borderline method.

The standard setting study is based on piloting of 178 reading items spread across 6 booklets. Each booklet contained two or three texts deemed suitable for learners with an A2 reading competence as well as two or three texts considered suitable for learners with a B1 reading competence. 1346 students of Norwegian as a second language answered two of the six booklets. The piloting took place two or three weeks before the students sat for either *Norskprøve 2* or *Norskprøve 3*.

In the study 1314 of the students (total 1346) are assigned to whole, or in-between, CEFR levels by their teachers. 178 reading items are assigned to whole or in-between levels by a team of 15 standard setting judges. 137 of these items survived the piloting.

#### **The Kaftandjjeva & Takala Compound Cumulative Method (The K&T method)**

- 1 A number of judges assess individual test items, in our case on the CEFR scale answering the question: *At what CEFR level can a test taker answer the following item correctly?*
- 2 Based on the aggregated judgments, the “core” difficulty bands for each level are established. Items are assigned to corresponding CEFR levels depending on the band in which the item difficulty falls.
- 3 Some items are reassigned to CEFR levels because the original assignment given by the judges does not match the empirical difficulty.
- 4 Cut-off scores are established on the basis of cumulative frequency distribution of items into CEFR levels.

#### **The Borderline Method**

- 1 Students with a competence on the border between two CEFR levels are focused. Teachers assign each of these students to an in-between-CEFR-level, for instance A1/A2, A2/B1 etc.
- 2 The mean test score for each in-between group is calculated. This mean test score becomes the cut-off score between two CEFR levels.
- 3 Some students are reassigned to CEFR levels because the original assignment given by the teacher does not match the established cut-off scores.

Since the main aim of the study was to assign items to CEFR levels, the next step was to find out how many per cent of the students in each CEFR group answered the different

items correctly. We decided that if more than 67% of the students in one CEFR group answered an item correctly, the item was assigned to the same CEFR level.

*Table 14.1 Example of level assignment of items based on the Borderline methods*

	P-values for items in different CEFR groups				
	A1	A2	B1	B2	Total
Item 10	.21	.33	.60	.96	.58
Item 11	.26	.53	.85	1.00	.74
Item 12	.87	.91	.98	.99	.95
Item 13	.87	.99	.99	1.00	.98

The table shows that item 10 is assigned to B2, item 11 to B1 and items 12 and 13 to A1.

Table 14.2 shows how many items are assigned to the same level by The Borderline method and the K&T method.

*Table 14.2 Number of items assigned to the same level by the K&T method and the Borderline method*

		Kaftandjieva & Takala c c method				Total
		A1	A2	B1	B2	1
Borderline method	A1	16	6	0	0	22
	A2	3	47	0	0	50
	B1	0	5	34	2	41
	B2	0	0	4	20	24
Total		19	58	38	22	137

The table shows that 117 of the 137 surviving items (85.4%) are assigned to the same levels by the Borderline method and the K&T method. The 81 items assigned to A2 and B1 we decided to treat as “core” A2 and B1 items and have a close look at them. Since the standard setting study is quite recent, we have just started the “scrutinising process”. Looking closer at the easiest and most difficult items, a pattern is discovered. Items are easy if:

- the text is structured in a way which make the correct answer easy to spot;
- the answer is located close to the beginning of the text;
- the same wording is used both in text and question, making the correct answer easy to spot;
- the answer can be copied directly for the text.

Items are difficult if:

- the test candidate have to make inferences;
- there is a lot of information to consider;
- there is information competing for the readers' attention;
- the vocabulary in general and/or single words makes things difficult to understand.

The characteristic features of easy and difficult items we discovered in our context, fit items at the fringes of, if not being clearly beyond, the levels the study is focussing on, i.e. below A2 for the easy items, and above B1 for the difficult. The next step will be to scrutinise the core A2 and B1 items and try to describe these in the contexts given.

## **5 Concluding remarks**

In the end discoveries made in connection with standard setting, may help us understand better what items targeting specific Framework levels “are” or what characterises such items. In addition they could add to the empirical foundation of the CEFR. Such information will hopefully also be helpful for standard setting judges and guide item writers to target specific CEFR levels better. In this way standard setting may add to the CEFR by strengthening the tie to second language learner data. At the same time it contributes to second language research. Knowledge about what is easy and difficult in connection with reading for instance, also tells us something about reading competence, what is acquired early/late in the language learning process, and what strong and weak learners are able to do.

Normally standard setting serves one function: to establish cut-off scores between different competence levels. In some contexts it may improve the validity of the cut-off scores if the link between items/tests and the CEFR is less opaque.

# References

Alderson, J. Charles (2007). The CEFR and the Need for More Research. *The Modern Language Journal*, 91/ 4, p 659-663.

Cizek, Gregory J. & Michael B. Bunch, M.B. (2007). *Standard Setting. A guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, London, New Dehli: Sage Publications.

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

Hulstijn, Jan J. (2007). The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency. *The Modern Language Journal* 91/4, p 663-667.

Jaeger, Richard M. (1989). Certification of student competence. In Linn, Robert L. (red), *Educational Measurement* (Third Edition) s. 485-511. Washington DC: American Council on Education, p 485-511.

Kaftandjieva, Felianka (2004). Standard setting. In *Reference Supplement B to the Preliminary Pilot version of the Manual for Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment*. Strasbourg, Council of Europe.

Kaftandjieva, Felianka. & Sauli Takala (2002). *Relating the Finnish Matriculation Examination English Test Results to the CEF Scales*. Paper presented at Helsinki Seminar on Linking Language Examinations to Common European Framework of Reference for Languages: Learning, Teaching, Assessment.

Moe, Eli. (2008). *Standard setting of A2 and B1 items*. Standard setting study 1/2008, Norsk språkttest, University of Bergen / Folkeuniversitetet, Norway. <http://fu.no/default.asp?avd=231&nyh=6702>.

Utdanningsdirektoratet (2006). Læreplaner – kunnskapsløftet. [http://www.utdir.no/templates/idor/TM\\_Tema.aspx?id=148](http://www.utdir.no/templates/idor/TM_Tema.aspx?id=148).

Vox (2005). Læreplan i norsk og samfunnskunnskap for voksne innvandrere. [www.vox.no/CommonPage.aspx?id=677](http://www.vox.no/CommonPage.aspx?id=677).

Weir, Cyril (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing* 22/3, p 1-20.









now you know



COUNCIL OF EUROPE  
CONSEIL DE L'EUROPE  
Language Policy Division  
Division des Politiques linguistiques



EUROPEAN ASSOCIATION  
FOR LANGUAGE TESTING  
AND ASSESSMENT

Cito, Institute for Educational  
Measurement

Council of Europe

European Association for Language Testing  
and Assessment (EALTA)

### Linking to the CEFR levels:

Research perspectives

Neus Figueras & José Noijons (eds.)

*'Th  
served  
mis*