Reference Supplement to the Manual for *Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment*

- **Contents**
- **Foreword**
- **Section A: *Overview of the Linking Process***

Language Policy Division
Division des Politiques linguistiques

CONTENTS

Foreword

The Language Policy Division of the Council of Europe in Strasbourg published in January 2009 a reference tool " Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) A Manual" and " Futher Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling" in order to assist member States, national and international providers of examinations in relating their certificates and diplomas to the *Common European Framework of Reference for Languages*.

This Reference publication accompanies the Manual. Its aim is to provide the users of the Manual with additional information which will help them in their efforts to relate their certificates and diplomas to the CEFR.

During the work on the Pilot Manual it was agreed that the Reference Supplement would contain three main components: *quantitative and qualitative considerations in relating certificates and diplomas to the CEFR and different approaches in standard setting.*

Dr. Norman Verhelst (member of the Authoring Group for the Manual), Dr. Jayanti Banerjee (Lancaster University) and the late Dr. Felianka Kaftandjieva (University of Sophia) undertook to write the various sections of the first edition of the Reference Supplement and Dr. Sauli Takala to edit the publication.. The authors have revised their contributions on the basis of comments from the editor. There have also been some comments from the other members of the Authoring Group and from the ad hoc advisory group. However, the authors have final responsibility for their texts. Dr Thomas Eckes and Dr Frank van der Schoot have written new sectiond for this revised edition

The authors' goal has been to try to make their contributions as readable as possible. They have avoided technical language (formulas, symbols etc) as far as possible and provided concrete examples, figures and tables to illustrate the exposition. However, demanding subject matter cannot be simplified beyond a certain point without risking oversimplification. Indeed, one of the authors' main concerns has been to caution about oversimplifications that many "rules of thumb" imply. The authors have, by contrast, tried to promote thoughtful application of various methods and approaches. With some effort, all persons working in language testing and assessment will be able to grasp the essentials and will have gained a deeper understanding of how to construct better tests and examinations and especially how to assess their quality. They will also be more aware of the complexities involved in relating certificates and diplomas to the CEFR.

Section A of the Reference Supplement provides a short overview of the linking process. This section is drawn from the Manual and is provided to help readers remind themselves of the approach proposed.

Dr. Felianka Kaftandjieva has written Section B on Standard setting. She has done considerable amount of work on standard setting specifically in relation to the CEFR. In Section B, the author notes that the link between language examinations and the Common European Framework for Language (CEFR) can be established in at least three different ways:

• direct linkage to the CEFR scales of language proficiency

• indirect linkage via linkage to some local scales of language proficiency which have already been linked to the CEFR scales

• indirect linkage via equation to an existing test already linked to the CEFR scales.

Whatever approach is adopted in the particular concrete situation, the author stresses that the linkage always requires standard setting and thus standard setting is a key element in the linkage process.

Section B underlines the potentially very high stakes of the examinations for the examinees, and seeks to promote better understanding by providing a review of the current status of standard setting, its theoretical framework and still unresolved issues. Section B does this by:

• giving a brief overview of the main trends in the development of standard setting methodology

• describing the major unresolved issues and controversial points

• discussing some of the major factors that affect standard setting decisions and their quality

• presenting some of the most common methods for standard setting

• outlining the validation process and providing evaluation criteria for the technical quality of the standard setting

• describing the main steps in standard setting procedures, and

• presenting some basic recommendations and guidelines for standard setting.

It will be obvious from the thorough review in Section B that there are several possible approaches for standard setting in relation to CEFR and the approach presented in the Manual is not the only appropriate one. Whatever approach is chosen, the validity of the claimed linkage depends on how well the various activities were carried out and how thoroughly and appropriately the results are reported.

Section C, written by Dr. Norman Verhelst, gives an overview of the main concepts and theoretical foundations of Classical Test Theory (CTT). Classical Test Theory has been used for more than fifty years as a guide for test constructors to understand the statistical properties of test scores, and to use these properties to optimise the quality of the test under construction in a number of ways. Section C reviews the main issues of Classical Test Theory and shows what can and cannot be expected from CTT. First, some basic concepts are presented followed by a discussion of procedures which are used in the framework of Classical Test Theory.

As the author's goal has been to make the text as accessible as possible for the non-technical reader, the first two sections (Basic Concepts and Procedures) do not contain any formulae. However, the author notes that as CTT is a statistical theory, it is not possible to present and discuss it in great depth without having recourse to the exact and compact mode of expression provided by mathematical formulae and, therefore, reference is made to formulae in a more technical section. These more technical sections are stand-alone elements, and follow the main text in the order they are referred to.

Section D, on qualitative analysis methods, is written by Dr. Jayanti Banerjee. The chapter provides an extensive overview of the range of qualitative methods available for investigating test quality. It demonstrates a large variety of options available and explains the key features of each, covering the following topics: an overview of qualitative methods, verbal reports, diary studies, discourse/conversation analysis, analysis of test language, data collection frameworks, task characteristic frameworks, questionnaires, checklists and interviews. In addition, examples of research using the methods are provided to illustrate how specific qualitative methods have been implemented.

The author suggests that many of the methods described could also be used as part of standard setting procedures and illustrates this in sub-section 6: Using qualitative methods in standard setting.
The author concludes that qualitative methods have considerable potential to explain and augment the statistical evidence we gather to assess test quality. Many of the methods are complementary and can be

used for the triangulation of data sources. The importance of the validity and generalizability of the data collection methods is stressed in order to legitimise the inferences drawn from them.

Section E, by Dr. Norman Verhelst, deals with Generalizability Theory and contains four parts. The first two parts give a non-technical introduction into generalizability theory. In the third and fourth sections the same problems are treated in a somewhat more technical way. The author notes that a very basic term of Classical Test Theory is not well defined: reference is made to repeated observations under 'similar' conditions, but 'similar' is not defined precisely.

A traditional way of controling for systematic effects is to try to standardize test administration as far as possible and feasible. Generalizability Theory was launched in the early 1970s to provide a method for assessing the effect of various factors on the measurement results. In the theory, measurements are described in terms of the conditions where they are observed. A set of conditions that belong together is called a facet. In this way, items and raters are facets of the measurement procedure.

Two important conditions in language testing are dealt with in more detail: the one-facet crossed design (persons by items) and the two-facet crossed design (persons by items by raters), and the possible application of Generalizability Theory in deciding on the optimal number of items and raters is demonstrated.

The author also discusses a problem which is commonly overlooked in using Generalizability Theory: typically every rater rates the same performances of the students to the task instead of every student generating an independent response for each rater. Yet, the design is treated as a two facet crossed design, which is not the case. This leads, in fact, to two different sources of measurement error: one attached to the student-task combination and one attached to the rater. This is a fundamental difference with the crossed model.

Section F, by Dr. Norman Verhelst, deals with a topic which has been a subject of discussion and debate in language testing for some time: *is language competence a unitary (unidimensional) or a multidimensional phenomenon*? If a test consists of several subtests, is it meaningful to report a single score or should test scores be reported separately for each subtest (in a profile)? Section F presents Factor Analysis - a well-established method (developed more than a hundred years ago) to test the dimensionality of the test in order to decide whether to report results using a single score or several scores. The author notes that although factor analysis was not defined originally as such, the model fits very well in the family of IRT-models discussed in Section G.

Section G, also by Dr Norman Verhelst, deals with the relatively more recent Item Response Theory (IRT). It consists of four non-technical sections (containing no formulae) where basic notions of IRT are explained and discussed. A number of notions and techniques are then discussed in a more formal and technical style. The author has strived to avoid the use of formulae as much as possible, making extensive use of graphical displays. To help the reader in constructing graphs using his/her own materials and using modern computer technology, a special section has been added with a step by step explanation of how most of the graphs in the section were produced.

Whereas the basic notion in Classical Test Theory is the true score (on a particular test), in Item Response Theory (IRT) the concept to be measured (in our case, language proficiency) is central in the approach. Basically, this concept is considered an unobservable or latent variable, which can be of a qualitative or a quantitative nature. If it is qualitative, persons belong to (unobserved) classes or types (of language proficiency); if it is quantitative, persons can be represented by numbers or points on a line. Only the latter case is dealt with in Section G.

One of the most attractive advantages of IRT is the possibility to carry out meaningful measurement in incomplete designs: it is possible to compare test takers with respect to some proficiency even if they did not all take the same test. This happens in Computer Adaptive Testing (CAT), where the items are selected during the process of test taking so as to fit optimally with the level of proficiency as currently estimated during test taking. Incomplete designs are also used in paper-and-pencil formats. Use of IRT methods requires a lot of technical know-how. This is sometimes packed in attractive software, and some users of this software may think that the problem is nothing more than technical know-how. The author warns that this is a naive way of thinking: the advantages of IRT are available if, and only if, the theoretical assumptions on which the theory is built are fulfilled. Therefore it is the responsibility of all users applying IRT to check these assumptions as carefully as possible. IRT methods are more powerful than methods based on classical test theory, but they may mistakenly be considered a methodology that ensures high quality assessment. The author, who has co-authored a very powerful IRT- programme called OPLM (One Parameter Logistic Model), warns against over-optimism which may be promoted by some enthusiastic proponents of IRT: " … using an IRT-model does not convert a bad test into a good one. A careless construction process cannot be compensated by a use of the Rasch model; on the contrary, the more carelessly the test is composed, the greater the risk that a thorough testing of the model assumptions will reveal the bad quality of the test." One practical consequence is that a separate assessment of the test reliability is always needed (preferably before IRT modeling) since it cannot be inferred from statistical tests of goodness-of-fit provided by software.

Section H, written by Dr. Thomas Eckes, deal with many-facet Rasch measurment. The chapter provides an informative introductory overview of many-facet Rasch measurement (MFRM). Broadly speaking, MFRM refers to a class of measurement models that extend the basic Rasch model by incorporating more variables (or facets) than the two that are typically included in a test (i.e., examinees and items), such as raters, scoring criteria, and tasks. Throughout the chapter, the author refers to a sample of rating data taken from a writing performance assessment and uses it to illustrate the rationale of the MFRM approach and to describe the general methodological steps typically involved. These steps refer to identifying facets that are likely to be relevant in a particular assessment context, specifying a measurement model that is suited to incorporate each of these facets, and applying the model in order to account for each facet in the best possible way. The author has chosen to focus on the rater facet and on ways to deal with the perennial problem of rater variability. More specifically, the MFRM analysis of the sample data is intended to illustrate how to measure the severity (or leniency) of raters, to assess the degree of rater consistency, to correct examinee scores for rater severity differences, to examine the functioning of the rating scale, and to detect potential interactions between facets. Relevant statistical indicators are successively introduced as the sample data analysis proceeds. In the final section the author deals with issues concerning the choice of an appropriate rating design to achieve the necessary connectedness in the data, the provision of feedback to raters, and applications of the MFRM approach to standard-setting procedures.

Section I of the reference supplement treats the Cito variation on the bookmark method in detail. This method of standard setting (presented in a more concise form in Chapter 6 of the Manual) is in some sense an ideal mixture of a student centered and an item centered method. The information collected on a (usually large) sample of student responses to the test items is summarized in a graphical way and is available to all panel members during the standard setting procedure. This releases the panel members from the difficult task of estimating success probabilities for borderline persons. The disadvantage, however, is that the panel must have a clear understanding of the essentials of Item Response Theory (IRT) to use the method effectively.

The author, Frank van der Schoot, for years the project director of the National Assessment Program for Basic Education in The Netherlands, has developed the method and applied it numerous times with teachers and teacher educators as panel members, all having many years of experience in teaching but with little or no training in psychometrics. This chapter of the reference supplement presents in detail how to explain

IRT to such panel members so that they can use the psychometric information correctly in the standard setting. I believe that the guideline provides an excellent basis for those who wish to implement the Cito variation of the bookmark method in their standard setting projects..

It is envisaged that new sections may be added to the Reference Supplement as new texts become available. One section was originally planned to deal with Test Equating, and hopefully that will materialise in the near future.

As editor of the Reference Supplement I am confident that it will prove very useful for the language testing and assessment community in general. It contains information which is not readily available in the mainstream language testing literature. More specifically it will provide good support for those who wish to contribute to the development of the Manual by providing feedback, by applying the Manual and by writing well-documented reports on some aspects or the whole process of linking examinations to the CEFR. I strongly believe that the Reference Supplement will also contribute to the improvement of language testing/assessment quality. Feedback and comments on the Reference Supplement (eg. suggestions for new sections) are invited. Please contact Johanna Panthier at Johanna.Panthier@coe.int

October 21, 2009
Sauli Takala

Section A: Overview of the linking process

The Manual for relating examinations to the *Common European Framework of Reference for Languages* (CEFR) presents four inter-related sets of procedures that users are advised to follow in order to design a linking scheme in terms of self-contained, manageable activities. All of the activities carried out in all four sets of procedures contribute to the validation process.

Familiarisation: a selection of activities designed to ensure that participants in the linking process have a detailed knowledge of the CEFR. This familiarisation stage is necessary at the start of both the Specification and the Standardisation procedures

In terms of validation, these procedures are an indispensable starting point. An account of the activities taken and the results obtained is an essential preliminary component of the validation report.

Specification: a self-audit of the coverage of the examination (content and tasks types) profiled in relation to the categories presented in CEFR Chapter 4 “Language use and the language learner” and CEFR Chapter 5 “The user/learner’s competences.” As well as serving a reporting function, this exercise also has a certain awareness-raising function that may assist in further improvement in the quality of the examination concerned.

These procedures assure that the definition and production of the test have been undertaken carefully, following good practice.

Standardisation: suggested procedures to facilitate the implementation of a common understanding of the “Common Reference Levels” presented in CEFR Chapter 3. Standardised exemplars will be provided to assist training in the standardisation of judgements.

These procedures assure that judgements taken in rating performances reflect the constructs described in the CEF, and that decisions about task and item difficulty are taken in a principled manner on the basis of evidence from pre-testing as well as expert judgement.

Empirical Validation: the collection and analysis of test data and ratings from assessments to provide evidence that both the examination itself and the linking to the CEFR are sound. Suggestions and criteria are provided for adequate and credible validation appropriate for different contexts.

These procedures assure that the claims formulated through Specification and Standardisation (“test-under-construction”) can indeed be confirmed when the examination is administered in practice (“test-in-action”) and data on how persons belonging to the target population behave when the test is so administered becomes available.

Relating examinations to the CEFR can best be seen as a process of "building an argument" based on a theoretical rationale. As noted above, the central concept within this process is "validity".

Evidently it is first necessary to ensure **Familiarisation** with the CEFR (Chapter 3) before linking can effectively be undertaken.

Then before an examination can be linked to an external framework like the CEFR (external validity), it must demonstrate the validity of the construct, and the consistency and stability of the examination (internal validity). To prove internal and external validity, quantitative and qualitative methods can be combined. Specification (Chapter 4) can be seen as a qualitative method: providing evidence through content-based arguments. The actions which result in filling in forms A1 and A3-A7 in Chapter 4 focus on the internal validity of the examinations. Forms A2 and A8-A20 focus in a qualitative way on the external validity. There are also quantitative methods for content validation but this Manual does not require their use.

Standardisation (Chapter 5) involves both qualitative and simple quantitative procedures - through training and comparison with calibrated test samples and performances - to prove external validity.
While the activities are mainly qualitative in orientation, quantitative evidence of the degree of success in the standardisation of judgements is also required.

Finally, **Empirical Validation** (Chapter 6) uses quantitative procedures based on data collection and analysis to demonstrate firstly "internal validity" and secondly "external validity". Chapter 6 demonstrates that proper empirical validation requires considerable psychometric know-how, just as test construction does. If such experience is not available to the examination providers, it is recommended that they arrange sufficient training or obtain the services of a qualified psychometrician.

The approach adopted in this process is an inclusive one. The recommended procedures in each of the chapters mentioned above encourage alignment of examinations to the CEFR with differing degrees of rigour appropriate to different testing contexts. The Manual aims to encourage the application of principles of best practice even in situations with modest resources and expertise available. First steps may be modest, but the aim is to help examination providers to work within a structure, so that later work can build on what has been done before, and a common structure may offer the possibility for institutions to more easily pool efforts in certain areas.

The recommended techniques are organised in a logical order in such a way that all users will be able to follow the same broad approach. Users are encouraged to start with Familiarisation and are guided through the options offered by the techniques for each of Specification, Standardisation and Empirical. They are asked to identify, from the range of techniques and options offered and similar techniques in the literature, those most appropriate and feasible for their context.

Not all examination providers may consider they can undertake studies in all of the areas outlined above. Some institutions in “low-stakes” contexts may decide to concentrate on specification and standardisation, and may not be able to take the process to its logical conclusion of full-scale empirical validation as outlined in internationally recognised codes and standards for testing and measurement. However, it is highly recommended that even less well-resourced examination providers should select techniques from all three areas. The linking of a qualification to the CEFR will be far stronger if the claims based on test specifications and their content are supported by both standardisation of judgements and empirical validation of test data. Every examination provider - even examination providers that have only limited resources or countries that have decentralised traditions - should be able to demonstrate in one way or another through a selection of techniques both the internal quality and validity of their examination and its external validity: the validity of the claimed relationship to the CEFR.

The different elements in the linking scheme outlined above are shown in Figure 1.1.

FIGURE 1.1: VISUAL REPRESENTATION OF PROCEDURES TO RELATE EXAMINATIONS TO THE CEF