Language Policy Division
Division des Politiques linguistiques

January 2009

Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)

Further Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling

Brian North (Eurocentres / EAQUALS)
Neil Jones (Cambridge Assessment / ALTE)

Language Policy Division, Strasbourg
www.coe.int/lang

Contents

1. Introduction **Page 1**

2. Constructing and Interpreting a Measurement Scale **Page 2**

2.1. Specification **Page 3**
2.2. Pretesting **Page 4**
2.3. Data Collection and Scale Construction **Page 4**
2.4. Scale Interpretation **Page 5**

2.4.1. Interpreting Cut-offs **Page 5**

3. Building an External Criterion into the Main Design **Page 6**

3.1. CEFR Anchor Items **Page 6**
3.2. Holistic Teacher Assessment **Page 7**

3.2.1. Assessment Instruments
3.2.2. Accuracy of Teacher Ratings

3.2.2.1. Interpretation of the Levels
3.2.2.2. Rating Invisible Skills
3.2.2.3. Lenience and Severity
3.2.2.4. Criterion- and Norm-referencing

3.2.3. Setting Cut-offs

3.3. Descriptors as IRT Items **Page 11**

3.3.1. Rating Scale
3.3.2. Teacher Assessment
3.3.3. Self-assessment
3.3.4. Setting Cut-offs

4. Exploiting the CEFR Descriptor Scale Directly **Page 14**

4.1. Benchmarking with FACETS **Page 15**

5. A Ranking Approach to Cross Language Standard Setting **Page 16**

6. Conclusion **Page 18**

1. Introduction

In setting standards, the issue of their maintenance over time, of their continuity within the developmental and operational testing cycle is of course fundamental. It is important to try and relate the setting and maintaining of standards to such a cycle. It is this area with which this document is essentially concerned.

It also suggests ways of using teacher judgements to set standards, using CEFR-based descriptors and/or assessment criteria to establish a link across languages. The document discusses approaches exploiting teacher and/or self-assessment to set the actual standards. But the fundamental emphasis is put on the need to see standard setting in the context of a scale of levels, a range of languages, developmental and operational cycles, and administrations over time. All these points concern scaling.

It is clear that in relating standard setting to the developmental and operational test cycle, over time the emphasis must shift from standard setting to standard maintaining. This does not mean that standards can be set once and for all. In developing a new exam it is very likely that the first standards set will be provisional and uncertain: an iterative cycle of progressive approximation is the norm. Even when the standard inspires confidence some procedure for longitudinal monitoring is necessary. None the less, the emphasis is on carrying a standard forward, not re-inventing it every session. This implies two things:

- Firstly, relevant approaches are comparative: this session's test and candidature are compared with previous tests and candidatures. This may provide a different focus for human judgment than the standard setting one.

- Secondly, relatively greater effort can and should be paid to the techniques that enable standards to be carried forward: that is, item banking and scaling.

The fact that scaling has become an increasingly important aid to standard setting is made clear in Chapter 6 of the Manual. Approaches to standard setting were presented in more or less chronological order in order to assist the reader in following the introduction of different concepts, and the user will have noticed certain trends over time:

a) More recent approaches – Bookmark, Basket, Body of Work – all consider standards in the context of the relevant proficiency continuum concerned: "It is B1 rather than A2 or B2", as opposed to "It is mastery"/"It is *not* mastery".

b) It has become standard procedure to feed information from pretest data to the panel, typically information on empirical item difficulty (for round two) and calculation of the impact the provisional decisions would have on the percentage of candidates passing (for round three).

c) IRT (Item Response Theory) is often used to place items from different (pre)tests onto the same measurement scale and to allow the cut-offs to be determined once on the item bank scale, rather than repeated for each new form of the test.

d) The most recent method described (Cito variation of the Bookmark method) not only encompasses the above three points, it also asks panellists to judge not the difficulty of the individual items, but the *cut-offs* between levels on the IRT measurement scale itself. It is thus a combination of a panel-based and a scalar approach.

This document follows up on this trend visible in Chapter 6 of the Manual, with the focus on constructing and using scales in the developmental and operational testing cycle. That is, it places item banking at the heart of standard setting.

2. Constructing and Interpreting a Measurement Scale

Space does not permit a detailed discussion of IRT. Baker (1997) offers a particularly good, simple introduction, and Section G of the Reference Supplement to the Manual describes IRT models in more detail. Here let us simply review the essential features of an IRT-based item banking approach, illustrated in Figure 1.

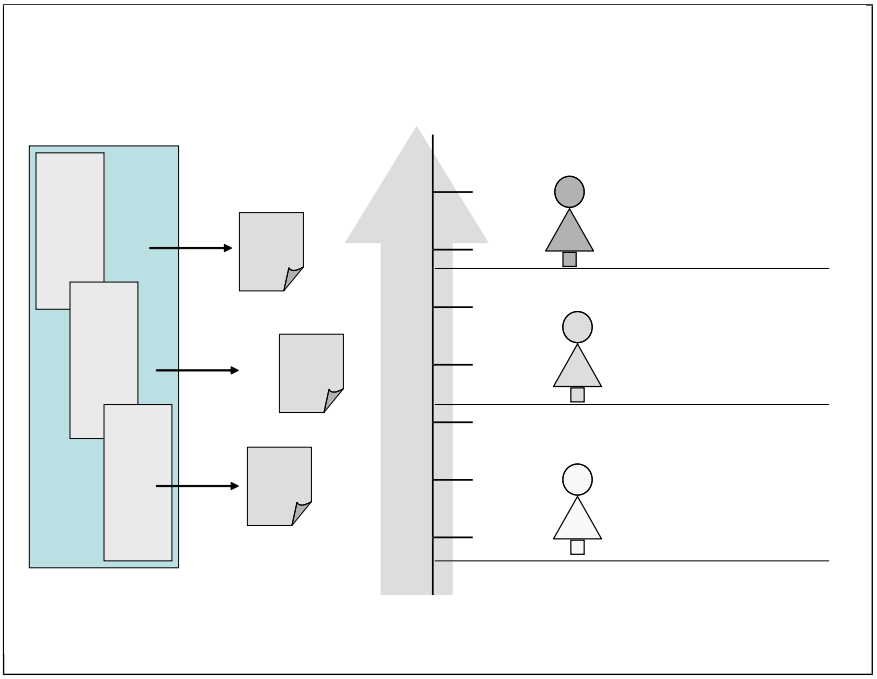

Figure 1: Item Banking in Outline

The approach provides a single measurement scale upon which we can locate items by their difficulty and learners by their ability, as well as criterion levels of performance. The scale is constructed from a possibly very large set of items. The vital thing is that the items are *calibrated* (their difficulty estimated) on a single scale, which is achieved by ensuring that all the response data used in calibration is linked up. This linkage can be achieved in several ways:

- tests with new material for calibration contain some items from the bank which are already calibrated (anchor items);
- a set of tests have common items, so all can be calibrated together;
- two or more different tests are given to the same group of learners.

When the scale has been constructed tests can be assembled using items at the appropriate level of difficulty for a target learner group. This ensures relatively efficient measurement. The learners' scores on tests can be translated into locations on the ability scale. While a learner's score differs depending on the difficulty of the test, the learner's location on the scale is an absolute measure of their ability, with a particular meaning. That meaning, of course, depends entirely on the items in the bank. "Ability" and "difficulty" are mutually defining – they arise in the interaction of learners with test tasks. What the scale measures is precisely described by the items and the way they line up by difficulty. It is by identifying the features of tasks which seem to make them more or less difficult that we can throw light on what exactly is being measured, and thus evaluate the validity of the item bank.

The item banking approach greatly facilitates setting and maintaining standards. Points on the scale may start off only as numbers, but, like the numbers on a thermometer, their meaning is constant, so that over time we can develop a shared understanding of the inferences and actions that they support. Interpreting performance on a non-item-banked test is a one-off: no understanding can be accumulated over time, and the inevitable uncertainty accompanying judgmental standard setting never diminishes. With item banking the fact that a

standard can be consistently applied facilitates progressively better understanding of what that standard means.

Item banking can thus be seen to support standard setting in a number of ways:

- Test forms at different levels, and parallel versions of a test at the same level, can be linked through an overlapping "missing data" data collection design.

- The initial batch of items calibrated can form the basis of an expanding item bank – still reporting results onto the same scale. Since an item bank has only one scale, CEFR cut-offs only have to be established once for the test scale(s) involved.

- Data from objective tests and from teacher or self-assessed judgments can be integrated into the same analysis; this facilitates:
  - involving a far larger number of language professionals in the interpretation of the CEFR levels that set the standards;
  - anchoring cut-offs across languages onto the same scale through the descriptors.

- An existing scale – perhaps reporting from a series of examinations at succeeding levels – can be related to the CEFR, without needing to suggest that the levels/grades reported match up exactly to CEFR levels. The pass standard for a test might actually be "between A2+ and B1; closer to B1 but not quite B1". A data-based, scalar approach enables this fact to be measured exactly and then reported in a comprehensible way (e.g. A2++), thus preserving the integrity both of the local standard and of the CEFR levels as insisted on by the 2007 Language Policy Forum (Council of Europe 2007: 14). One must not forget that, as stated in Manual Chapter 1 and emphasised in the foreword to the CEFR, the CEFR is a metalanguage provided to encourage reflexion and communication; it is not a harmonisation project telling people what their objectives should be (Council of Europe 2001: xi Note to the User).

The exact approach taken to developing an item bank will depend very much on the starting point, the purpose and the feasible scope of the project. Item banking can be applied to a suite of tests covering a range of levels, or to a single test at one level. The first of these situations is the more challenging because of the need to create good links across levels (*vertical* links). If there is the possibility of re-using items over time this makes item banking a particularly attractive option and enables better calibration. One may be starting from scratch or from existing test material, working with one skill or several, or even with the aim of constructing a framework for several languages. Curricular constraints may apply. It may be more or less feasible to run pretests or set up ad-hoc data collections. One may have more or less easy access to item response data from live exam administrations. Levels of technical support, e.g. with analysis or with constructing a system to hold the item bank, may vary. These factors can be decisive in determining the exact approach adopted – and indeed the feasibility of an item banking approach at all. Adopting such an item banking approach may also require changes to the current practice in test development.

However, despite this variation in the design of specific projects, the following four basic steps are involved in creating a scale of items linked to the CEFR:

- Specification;
- Pretesting;
- Data Collection and Scale Construction;
- Scale Interpretation.

2.1. Specification

Whether one is working with an existing test or developing a new one, it is necessary to demonstrate how the test content relates to the CEFR.

The first step is therefore to create a very detailed CEFR-related specification of the skill(s) concerned at each level. In so doing, the specification procedures outlined in Manual Chapter 4, and the Content Analysis Grids there referred to could be useful.

To develop new test material a team needs to be selected. It is vital to carry out familiarisation and standardisation training with them as outlined in Manual Chapters 3 and 5 in order to ensure that they share the consensus interpretation of the CEFR levels. Individuals or sub-teams are then assigned to prepare tests or pools of items targeted at relevant levels in which they have expertise.

2.2. Pretesting

The next step is that sets of items should be pretested on small, representative samples of candidates at the appropriate level(s). In practice it may be difficult to develop an item bank as a pure research project outside an operational exam cycle. Item banking is difficult to operationalise in the context of an existing examination if there is no pretesting in the normal test construction cycle. It may be necessary to introduce such a step if it is not already current practice.

If one exploits pretesting, then of course there are a number of issues that need to be considered as discussed in detail in Section 7.2.3 of the Manual. The most obvious is the security of items: whether learners engaged in pretesting are likely to see the same items in a live exam.

2.3. Data Collection and Scale Construction

The next step is to organise the items for initial calibration into a series of linked tests. The most practical method may be to include an anchor test common to all forms at a given level. More complex links are possible but may complicate the logistics of data collection.

Cross-level links need particular care, because they are difficult to build into an operational cycle, so require specific organisation. Picking the right target group is tricky; because of the possibility of targeting error, it is better to go for a relatively wide range of ability and be prepared to reject off-target responses (with very high or low facility). Generally vertical linking will work better if it can be done subsequently to calibrating items at each level, using items hand-picked for their difficulty and statistical performance (fit).

Calibrating a set of items covering a range of levels might theoretically be done in a single analysis containing all the response data. However, even where possible this process should be undertaken critically and iteratively, cleaning the data to ensure the most plausible result. If this approach is adopted, then the safest method is that applied by Cito: anchor each test 50% upwards and 50% downwards to its adjacent tests in such a way that every item is an anchor item, except 50% of the items on the highest and on the lowest test forms.

In practice the analysis and calibration involved in setting up an item bank is more likely to involve a number of iterations and extend over a longer period, and once integrated into the operational testing cycle, of course, becomes an ongoing process.

However the data for calibration are collected, the basic rules for quality control remain the same:

- Try for a sample of candidates reasonably representative of the live population.
- Try for an adequate sample size (say, 100 if using the Rasch model).
- Avoid off-target responses, that is, very high or low scores; remove them from calibration data where they occur.
- Try to avoid effects that will cause predictable bias, e.g. differential effort between an anchor test and a live test, or time pressure effects that make late items appear more difficult.
- Where such effects may be present, try to detect them and remove them from the calibration data.
- Try to pick anchor items carefully. Good anchors have *average* fit and discrimination indices and don’t function differentially across major groups of interest.

2.4. Scale Interpretation

This section relates primarily to the interpretation of a scale covering a range of levels (e.g. A2–C1), but it also has relevance to a scale targeting a single level (e.g. B1, so A2+ to B1+ in practice). The interpretation is portrayed as a process completed at a single sitting, although as pointed out above, in practice the process of completing the construction and interpretation of a scale may well be an extended and also an iterative business.

The first step in interpreting the scale is to identify the core space taken up on the scale by the items targeted at each level – without overlap to items at the next level, comparing items to the specifications in detail in this process. A decision then needs to be taken as where to place the provisional cut-offs within the area of the overlap. There are different ways of doing this and ideally decisions should be made in an iterative process taking account of all three. If the items were well written and well targeted, then these three perspectives should converge:

- Match each item in the area of overlap to the detailed specifications and CEFR descriptors for the adjacent levels concerned, to inform judgment.

- Place the cut-point exactly in the middle of the overlap between the items originally targeted at different levels, as suggested in the discussion of the Bookmark method and Cito variant of it in Manual Sections 6.8. and 6.9.).

- Set the cut-off within each area of overlap in such a way that the space taken up on the logit scale by each level makes sense in proportional terms. That is, levels shouldn't differ haphazardly in width.

During this process, it is valuable to investigate and seek explanations for items that are "outliers" from a test (items designed for one level that have actually landed in another one). In some cases a mismatch between text and item(s) may be the cause; in other cases the task may seem reasonable and well calibrated and just turn out to be an easier or more difficult challenge in the microskill that is still a valid objective at the adjacent level. A decision needs to be made whether to keep or discard each such item. Generally speaking items whose intended difficulty and empirical difficulty differ greatly are suspect.

In this process, the decisions might be taken by a designer/analyst alone, but there is strength in numbers. It is possible that a small panel would produce more representative results. Certainly the exact process followed should be documented and reported.

The reader will have noticed that the procedure described in this section has similarities to the Item-descriptor Matching method (Manual Section 6.7.) and the Bookmark method (Manual Sections 6.8 –9), and the Cito variant of it (Manual Section 6.9.). The difference is in who determines the level of the items and when. As described here, it is the item writers who construct items to a careful specification, informing a quasi-mechanical setting of cut-offs. In standard setting procedure such the Item-descriptor Matching method and the Bookmark method it is a panel of judges who do something similar as a post-hoc exercise. In both cases one would like to say that some additional external validation is still highly desirable. The experience of serious language testing operations, who invest much effort in the item writing process, is that skilled item writers can produce valid tests which are well targeted at the intended level. But this alone is too imprecise to support vertical equating or accurate grading. For that one needs an IRT model, scaling and a known standard (such as a CEFR level).

2.4.1. Interpreting Cut-offs

The procedure described in Section 2.4. lets one set a cut-off on an item bank scale that divides as cleanly as possible the items intended to describe two adjacent levels. But this is not yet an appropriate cut-off for saying that a candidate has achieved the level. A candidate at exactly the level of an item – say, the first and easiest B1 item – has just a 50% chance of succeeding at that item. For each more difficult item the

probability is even lower. This borderline candidate's score on a test made up of items appearing between the A2/B1 and the B1/B2 cut-offs on the scale would be rather low – too low to imply achievement of the level. So as explained in Manual Section 6.8., we must decide on a response probability (RP) which is considerably higher than 50% – say, 80% – and set the cut-off higher, such that a candidate at the cut-off would have an 80% chance of succeeding on the easiest B1 item, and a reasonable score on a test consisting of B1 items.

Looked at another way: the A2/B1 cut-off is also the top end of Level A2. The borderline B1 candidate has not just achieved but fully mastered A2, with an 80% chance of succeeding on the most difficult A2 item and an even higher chance on easier items.

3. Building an External Criterion into the Main Design

Section 2 outlined classic steps necessary to develop any scale: (a) define the construct and prepare the items; (b) check them through pretesting; (c) collect data and create a scale, and then (d) interpret the result and set provisional cut-offs. However, one would still like to do more to verify that the CEFR cut-offs so set are in fact accurate. The obvious method of corroborating cut-offs arrived at in this way is to undertake a cross validation study: i.e. use another standard setting technique. This could be one of the techniques outlined in Manual Chapter 6, or under the heading of *external validation* in Manual Section 7.5. – for example, the worked example given there of using teacher judgments as an external criterion.

A scalar approach offers the possibility of building such external criteria into the data collection design itself. If items from an existing CEFR test/scale are included, or if a candidate-centred approach is adopted with teacher judgments or even self-assessments, then these external perspectives can be taken into account in setting the initial standards. It is certainly not the case that standard setting and external or cross-validation need to be separated in time. Integrating the two into the initial data collection is greatly to be recommended, since it can ensure that the linking study remains "on track" throughout and can so avoid the danger of a contradictory result from a conventional external validation study.

There are at least three ways in which the external criterion can be built into the main project design. It is a good idea to include more than one approach, since the more information for standard setting that can be included the better.

- **CEFR Anchor Items:** Integrate the CEFR illustrative items for the languages concerned into the data collection test forms.

- **Holistic Teacher Assessment:** Involve the teachers of the candidates concerned, asking them to give assessments of CEFR level for the skill(s) concerned, guided by training and an appropriate CEFR-based rating instrument.

- **Descriptors as Items:** Use appropriate individual CEFR descriptors as separate items on a checklist for teacher assessment and/or for self-assessment.

If teachers are involved in the project in this way, then this opens the possibility of holding standard setting seminars with some of them (using a method such as those in Manual Chapter 6) in order to ensure that the items for the data collection are indeed well targeted at the typical range of levels of their students.

3.1. CEFR Anchor Items

Including CEFR illustrative test items for the skill(s) concerned is an obvious step to take. It can be done in two ways, at different phases of the project. The two approaches are not mutually exclusive; an ideal project would include both:

- **Pretesting:** In pretesting, CEFR illustrative items can be included in order to check that assumptions being made about the match between local items and the CEFR levels are broadly correct, and that the local items for a level (a) cover a similar range of logits on the scale to the illustrative items and

(b) show comparability of means.

- **Data collection:** In the main data collection for an item banking design. Here, these CEFR illustrative items could form part of the batch of anchors linking adjacent tests, thus helping to form the actual scale. If exploiting a live exam to collect item response data for calibration, one must think how that CEFR anchor data can be linked in. One possibility is to require candidates to complete an anchor test of calibrated items at about the same time they take the live exam; however, a likely problem is that the anchor test may be taken less seriously than the live exam, leading to a predictable bias effect.

Szabo (2007) demonstrates the usefulness of integrating CEFR illustrative items. He found that, even though the CEFR illustrative items available at the time were far from perfect from various perspectives, he was able to exploit them as reference items to demonstrate that there were no significant differences in the spread of difficulty on the scale covered by the new items and the illustrative items.

3.2. Holistic Teacher Assessment

Teachers can be a valuable source of information in terms of both collateral information on the candidates and in terms of contributing assessments onto the CEFR levels that can be exploited in setting cut-offs on the measurement scale.

Jones, Ashton and Walker (2007) provide an example of teacher ratings playing an important part in setting cut-offs and contributing to vertical scale construction, from the English "Asset Languages" project. This project reports assessments in terms of grades on a CEFR-based "Languages Ladder". Teacher ratings were collected at two stages:

- **Pretesting:** At the pretesting stage, with a global rating elicited by reference to English National Curriculum levels, which are broadly in line with the grades on the Languages Ladder. These National Curriculum levels were found to be the most familiar proficiency levels for primary and secondary school teachers to use.

  At the pretesting stage teachers can give a single, global rating of level for each of the candidates. These ratings corresponded to ability on a pre-defined scale, so that the IRT analysis of pretest data can be anchored to the scale via the estimated ability of candidates.

- **Data Collection:** In the main data collection, with ratings for listening and for reading, elicited by reference to the Languages Ladder/CEFR, using an instrument drawing on both scales.

  When live tests are constructed, the item difficulties anchored in this way can be used to estimate the ability corresponding to different scores in the test, and hence the score corresponding to each grade threshold on the scale.

The pre-defined scale is a template which applies by default to all languages and objectively tested skills (reading, listening) in the Asset Languages framework. It represents an expectation of what the scale should look like, proportionally, when empirically constructed. It is related to the common scale which underlies Cambridge Assessment's ESOL exams, itself empirically constructed and interpretable as showing that levels emerge as a function of learning effort and observable learning gains (Jones, Ashton and Walker 2007). Over time an empirical vertical linking should enable refinement of each scale, but for languages where this is yet to be achieved, it is the teacher ratings which effectively anchor the levels.

Asset Languages works with 25 languages defined at six levels. Aligning such a large number of elements in a complex framework is beyond the capabilities of conventional test-centred approaches to standard setting. In addition, the requirement of internal coherence, it can be argued, is best met by working from a model of the framework as a whole downwards to the elements within it, rather than working upwards from unrelated standard setting decisions concerning individual languages, levels and items.

As well as being used to help to set the standards as discussed in this section, holistic assessments by teachers can also be used in an external validation study to corroborate cut-offs set by another method. North (2000b), for example, reports using teacher holistic assessments of "knowledge of the language system" and of writing to corroborate provisional item banks of grammar and vocabulary items for different languages in the context of an intensive language learning course in a country where the language concerned is spoken. This method is described in Manual Section 7.5.3. whilst discussing external validation.

3.2.1. Assessment Instruments

There are different forms that holistic teacher assessments could take:

a) **Holistic Scale:** A single holistic judgment answering the question "What level is this person," using whichever of the CEFR scales and subscales is most appropriate, or the simple holistic scale given as Appendix Table C1 in the Manual. A single item "test".

b) **Analytic Grid:** A number of judgments in relation to CEFR scales for different communicative language activities and aspects of communicative language competence. One could use rating instruments developed from CEFR scales themselves, or an analytic descriptor grid derived from them such as those in Appendix C to the Manual (Tables C2–C4). This would at least give a "test" of five or six items. The result could of course then be averaged to give a single "global result", matching the single holistic judgment above.

c) **Checklist:** Alternately, judgments on checklists of 30–50 descriptors could be required. Such checklists were used in the Swiss research project that calibrated the descriptors and developed the CEFR levels and nowadays are commonly available in the Language Biography section of validated European Language Portfolios (ELP). This approach could yield "tests" of 30–50 items. One way of using the scores is to allocate a single CEFR rating based on the total number of "Can Do" descriptors endorsed. What percentage endorsement would demonstrate achievement of the level in question? There is no simple answer to this question. European Language Portfolios generally adopt 80%. However, provided the selection of descriptors is reasonably representative, if a learner "can do" what is stated on 67% of the descriptors, they could be considered the level concerned. If they achieve 85%, or 90% then they are probably well into the next level, even though they will not yet have fully mastered it; in other words, given a checklist for the succeeding level, they would probably already say they "can do" many of the descriptors.

   Another way of using such data is to treat each descriptor as a separate test item, as discussed in the next section.

3.2.2. Accuracy of Teacher Ratings

Whichever approach to teacher assessment is used (scale, grid or checklist) it is good practice to consider the assessment as a test procedure that needs to demonstrate a certain validity. Aspects to bear in mind include the following:

- whether teachers understand the levels;
- whether teachers are actually able to judge the skills in question;
- whether differences of leniency/severity will affect the standard setting;
- whether teachers assess according to the criteria, or norm-reference between stronger and weaker learners.

3.2.2.1. Interpretation of the Levels

The teachers must be trained as well as possible – ideally through the Familiarisation and Standardisation processes described in Manual Chapters 3 and 5. However, systematic training for all teachers involved in the project may be difficult to achieve, as the collaboration of teachers, who already have more than enough to do, is often dependent on their good will. In addition, finding dates for training sessions may be difficult – and some teachers will then inevitably, for one reason or another, be unable to attend. Distance training (e.g. studying the examples on CEFTrain.net) may be prescribed, but a system to check whether the teachers actually do their homework will be necessary.

One way of dealing with this problem would be to flag in the data whether teachers completed training or not, with the option to reduce to a more reliable group of respondents, should difficulties occur in the data.

3.2.2.2. Rating Invisible Skills

The second issue is whether teachers are actually able to give accurate judgments on the skills in question. Clearly it is the performance skills of speaking and writing which are uppermost in our perception of a learner's level. Reading and listening being mental processes, only indirectly observable, it is much harder even for a teacher who knows their pupils well to rate them accurately.

It is predictable that the ratings teachers provide should be coloured most strongly by the performance skills. How much of a problem this is for using teacher ratings is hard to say. It would be no problem at all if we could count on a population of learners having a "flat profile" across skills, but individuals, as we know, rarely do have a flat profile – they are better in some skills than others. The assumption of flat profiles (rather than much higher competence in receptive skills) would certainly be a very questionable assumption in many contexts, for example, assessment of English ability in Northern European countries. Grin (1999/2000) demonstrates that well over 20% of the English language competence of Swiss-German speakers cannot be accounted for by study, work, relationships or travel; it is just "in the air" – and probably more of it is receptive than productive[1].

Thus in eliciting holistic judgments by teachers, a key issue is which skills to focus on:

- the productive skills, speaking and writing;
- "global CEFR level";
- impression judgments on the candidate's (invisible) ability at the skills under study (listening, reading, linguistic knowledge).

It is difficult to know in advance whether teachers will be capable of making consistent judgments about receptive skills that correlate highly enough with test scores to be useable. They may instead actually give a global rating coloured more from what they see of student performance (in the productive skills) than give a genuine rating of the receptive skill(s). Asking for a profile for each candidate ensures that the teacher at least thinks about the issue. The assessments with the highest correlation to the test results could be then actually exploited in the standard setting.

Even if one cannot use ratings of productive skills as proxies for the receptive skills, it seems a good idea to collect such data in a project of this kind. After all, the logistical problems with this type of project are more concerned with the scope (number of candidates, teachers, administrations, etc.) and the necessary familiarisation and standardisation training of those involved (item writers, test developers, teacher-raters). Adding one or two extra aspects for each teacher to rate for each candidate, so that they give a profile of four or five aspects, is a small increment logistically for teacher and analyst – but could have a pay off in the standard setting.

[1] By contrast, the French language competence of Swiss-German speakers and the English and German competence of Swiss-French speakers are almost entirely explained by those four factors in Grin's data.

3.2.2.3. Lenience and Severity

Raters are known to vary in strictness. How then can one feel confident that differences in leniency/severity between teachers do not cause large unwanted parts of the variance? If teachers rate only their own pupils there is no way of checking this, although it is reasonable to assume that the effect will average out across teachers. If teachers can be given the opportunity to rate a number of the same scripts or DVD samples, then there is a possibility of using the many-faceted Rasch model as operationalised in the program FACETS (Linacre 1989; 2008) to detect rater severity and take account of it in estimating ability estimates for the test takers. This will still not guarantee that the teachers overall are on target: the only way to do this would be to use the ratings of one or more authoritative experts to anchor the analysis.

3.2.2.4. Criterion- and Norm-referencing

When teachers are asked to rate a range of learners in their class, there is a tendency for them to exaggerate the differences between their stronger and weaker learners. They want to give credit to better learners and tend to dislike giving two learners the same grade (which might be required in strictly following the criteria) when they can clearly see a difference in the achievement of the two learners. There was evidence of this with the Swiss teachers involved in the project to calibrate the CEFR descriptors that was remarkably similar in the two separate data sets for the two years. The solution adopted was to exclude the data from the top and bottom learner from each class (theoretically at the 25th and 75th percentile) and to exclude two more learners from teachers showing a spread of judgments more than two standard deviations from the mean. (North 2000a: 215–6).

3.2.3. Setting Cut-offs

The next issue is the question of how such holistic assessments of candidate level (one result per candidate, e.g. “B1”) can be related to scores on a test or points on an IRT scale. Classic methods for relating raw scores to the teacher assessments, the Contrasting Groups and Borderline Group methods, have already been discussed in detail in Manual Section 6.5. These are most useful when relating results from one test to one standard, but can also be adapted to relate test results to a series of ascending standards, like the CEFR levels.

Displaying data graphically is a good way of exploring what they can tell you. The following example uses “box plots” to show how scores from a test relate to several CEFR levels. Figure 2 is based on an artificial data set, containing the test scores of 750 candidates on a 50 item test. The CEFR level for each candidate comes from teacher assessment, using three adjacent levels. These levels are labelled as 1, 2 and 3 along the x-axis of the figure, but could correspond to A2, B1, B2, for example. The y-axis represents the scores on the test[2].

The three graphics depicted represent groups of candidates who were rated Level 1, Level 2 and Level 3 respectively by their teachers. The shaded box represents the middle 50% of the scores; for Level 1, these central scores range from about 8 to 15 on the 50 item test. The horizontal line across the shaded box (shown going through the dot) corresponds to the median. The upper border of the box shows the 75th percentile and the lower border represents the 25 percentile. Above and below the box itself are the so-called “whiskers” representing outlying scores. In this case the upper border of the whiskers corresponds to the highest score, and the lower border to the lowest score. Sometimes, the 10th and 90th percentile are displayed, together with the exact position of the outliers.

The plot is very clear in this case: the test discriminates quite well between the three levels, although there does remain some considerable overlap in the scores – which is not unusual. In order to relate holistic assessments to an IRT scale reported from a test or from a series of tests in an item bank reporting results onto the same scale, then logistic regression, explained in Manual Section 6.6.2. is a more appropriate technique. Most analysis packages have a function for plotting such logistic regression. In the “Asset

[2] The section on box plots is taken from the pilot version of the Manual, and was originally written by Norman Verhelst.

Languages" project described above, IRT methods were used to impute pupil abilities from teacher ratings, providing the basis for anchoring item calibrations and subsequently generating grades on a pre-defined proficiency scale for each test.

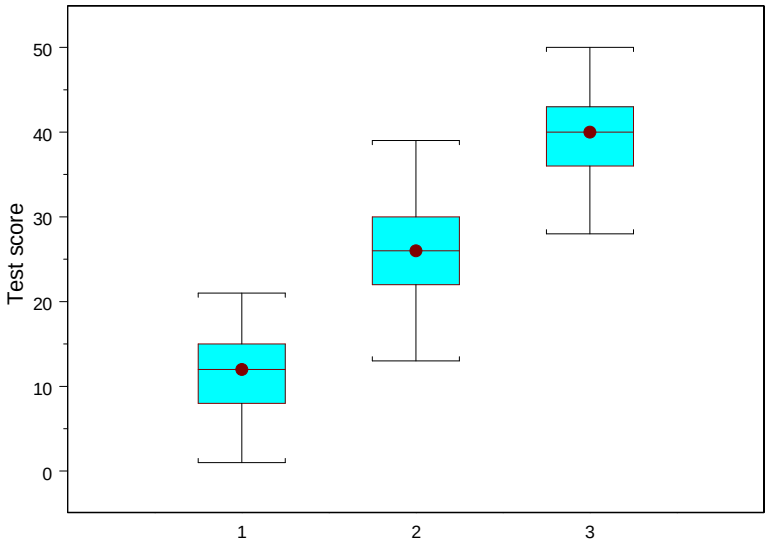

Figure 2: Box plot

3.3. Descriptors as IRT Items

When using descriptors on checklists, another approach is to treat each descriptor as a separate test item for teacher and/or self-assessment, as opposed to converting scores on checklists to a holistic result as suggested in Section 3.2.1.(c). The checklist of descriptors thus gives 30–50 items, analysed separately.

The exploitation of CEFR descriptors in data collection in this way has the following advantages:

- **Data-based standard setting.** It enables a large-scale data-based form of candidate-centred analysis; this can be exploited to base the standard setting on a wide consensus of the interpretation of the levels. This can be seen as a complementary view to that of a standard setting panel. The claim to empirical validity of the CEFR descriptors themselves is based on the fact that they reflect the consensus of a large group of practitioners rather than that of a single committee.

- **Cross-linguistic anchoring.** It can provide a means of anchoring different languages and regions into a common frame of reference, to the extent that we are satisfied that they are interpreted similarly. In the Asset Languages project, learners taking tests in two different languages provided self-ratings for each language in order to evaluate the potential of this approach as a way of verifying cross-language alignments of objective tests. This is an attractive idea because self-ratings by plurilingual informants should reflect reasonably well their relative competence in the two languages. The tendency of individuals' self-ratings to vary in terms of absolute level should thus not be a problem. The study replicates a design previously adopted in checking standards for the BULATS test across languages, by exploiting self-assessments for each language with ALTE "Can Do" statements.

- **Cyclical standard setting.** Use of descriptor-assessment by teachers and/or learners in each phase of a project (pilot, pretest, data collection) can encourage a cyclical approach to standard

setting, focusing on the issue from different angles (judgmental/data-based) that can ensure a successful project that stays "on track".

The potential weakness is that showing correlation by itself does not actually *build* a validity argument. Such a correlation between test scores and ratings does not demonstrate construct validity. Such a correlation could, in theory, be due primarily to a different construct (e.g. age). Hypothetically speaking, it would therefore in fact be possible to "relate" maths tests for 8–14 year olds to the CEFR levels in this way – through the construct "age" rather than "communicative language competences". This means that a strong specification argument must be built, following the procedures such as those outlined in Manual Chapter 4, before undertaking standard setting using descriptors. But self-ratings can themselves also be a valuable tool for construction validation. For example, Ashton (2008) used self-ratings by A1–B1 secondary school readers of German, Japanese and Urdu in a mixed-methods approach that compared their perception of the construct of reading at each level. The analysis produced interesting evidence of similarities and differences across languages and proficiency levels.

3.3.1. Rating Scale

The descriptor-items can be scored with a binary Yes/No or with a rating scale (0–2, 0–3 or 0–4). For example, the scale used in the Swiss CEFR research project was:

0	**1**	**2**	**3**	**4**
Describes a level *beyond* his/her capabilities	**Yes, in favourable circumstances**	**Yes, in normal circumstances**	**Yes, even in difficult circumstances**	**Clearly better than this**

For speaking, this was elaborated in instructions as follows:

> This describes a level which is definitely *beyond* his/her capabilities. Could *not* be expected to perform like this.
> Could be expected to perform like this provided that circumstances are favourable, for example if he/she has some time to think about what to say, or the interlocutor is tolerant and prepared to help out.
> Could be expected to perform like this without support in normal circumstances.
> Could be expected to perform like this even in difficult circumstances, for example when in a surprising situation or when talking to a less co-operative interlocutor.
> This describes a performance which is *clearly below* his/her level. Could perform better than this.

3.3.2. Teacher Assessment

In pursuing such an approach, it is important not to demand more of teachers than they can provide. Asking a teacher to complete a 0-4 rating scale for each "Can Do" descriptor on a 50 item questionnaire for every pupil in a class of 30 is probably not realistic and will at best produce poor data.

The approach used by the Asset Languages project was to ask teachers to complete a questionnaire for a "good", "average" and "weak" pupil in their class, and also to provide a ranking of the pupils showing where they fell in this classification. A similar approach was used in the Swiss project: teachers ranked the learners in each of two classes and then selected five learners from each class: the median learner; the learners at the 25th and 75th percentile, and the learners at the mid-point between the median learner and those two learners.

3.3.3. Self-assessment

Self-assessments with checklists of descriptors have been previously used to relate tests to proficiency scales (TOEIC/TOEFL: Boldt et al 1992; Wilson 1989, 1999, 2001; ALTE: Jones 2002). Checklists could be made with CEFR or ELP descriptors. The advantage of using CEFR descriptors is that they are largely based on descriptor-items which have been calibrated. The scale of descriptors with calibrations from the Swiss project is given in the appendix to North (2000a).

Opinions vary about the reliability of self-assessment in a standard setting project; however, many projects have used self-assessment data in this way. The problem if there is one is not with the calibration of the descriptors, because candidates tend to perceive their relative difficulty in similar ways. The difficulty parameters of descriptors estimated from teacher assessment and self-assessment therefore will show a high correlation. As an example, in the CEFR research project the correlation between the scale values of the descriptors derived independently from teacher assessment and from self-assessment correlated 0.98. It is candidates' assessment of their own proficiency that is less reliable: they tend to over- or under-estimate themselves.

It has often been observed that low-level learners tend to over-estimate themselves, while high-level learners tend to under-estimate. Intuitively pleasing explanations can be given: low level learners are elated with what they do know, and don't realise how much there is still to learn; high-level learners by contrast have come to realise that there is always more to learn. Actually it is not necessary to appeal to such theories to explain what we observe. Self-assessment is inherently error-prone, and both groups will tend to err in one direction: upwards in the case of low-level learners and downwards in the case of high-level learners. It is a predictable complicating factor in interpreting such data.

Despite the individual variation in the self-ratings for the reasons discussed above, Jones (2002) reporting on the ALTE "Can Do" Project found a strong correlation between self-rating and the proficiency level reported by Cambridge ESOL exam results, when data were grouped by exam. Four hundred and seventy-eight self-ratings on ALTE "Can Do" descriptors and on a sub-set of particularly stable CEFR descriptors for aspects of Fluency were linked to exam grades in ESOL exams over five levels.

Figure 3 shows the mean self-rating of candidates grouped by the exam grade which they achieved. Such data appear to enable quite detailed interpretation, suggesting for example that an 'A' grade at FCE corresponds approximately to a "C" pass at the next level up (a relation corroborated by other evidence).

Where there are doubts about the absolute level of self-assessment data, it can best be anchored by comparing it with teacher ratings.

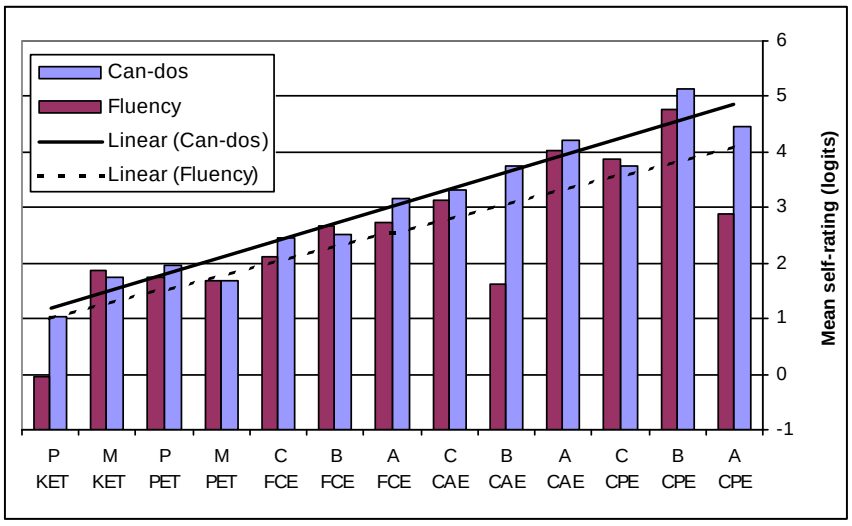

Figure 3: Mean Self-ratings by Exam Grade (ALTE "Can Do" Project)

3.3.4. Setting Cut-offs

Apart from summing results into a single result and giving a holistic rating of CEFR level as discussed above in Section 3.2.1.(c), there are at least three ways in which data from teacher and/or self-assessments in relation to checklists of CEFR "Can Do" descriptors can be exploited in an IRT analysis of data from both objectively scored test items and subjectively scored descriptors, in order to set standards (= cut-offs between levels) on the theta scale produced from the analysis.

a) The position at which the descriptors for different CEFR levels are placed on the scale can be used to make decisions about where one level stops and the next level starts. This approach relies upon easily seeing what level the CEFR descriptors are. For this purpose, it would be a good idea to code the descriptors for the analysis is such a way that the level is clearly visible. For example, an A2 descriptor: "*Can get simple information about travel*" might be coded "*A2-Travinfo*". The coded descriptors will then indicate visually the range on the scale covered by the descriptors for each CEFR level.

   The approach taken by the analyst – or panel – is a sort of assisted variation of the Bookmark method described in Manual Section 6.8; there is simply strong guidance for deciding where to put the bookmark. However, there may well be some overlap between descriptors for different levels. Judgment will therefore be necessary in order to decide the exact cut-off. This involves balancing the factors outlined in Section 2.4.

b) The position at which the descriptors for different CEFR levels are placed on the scale can be used as an additional step in a panel-based standard setting procedure, as a form of external validation.

   This approach is described in Manual Section 7.5.4.2. whilst discussing external validation, with the Cito variation of the Bookmark method as an example.

c) The CEFR research project descriptors can be used to exploit directly the cut-offs on the measurement scale underlying the CEFR descriptor scales. In this way, cut-offs are linked rather than items.

   This approach is discussed below in Section 4.

4. Exploiting the CEFR Descriptor Scale Directly

If CEFR descriptors are used as separate items for an IRT analysis, then there is the possibility of transforming the result of the teacher or self-assessments obtained in the new study into difficulty values on the same scale as the original CEFR research study.

One advantage of this approach is that the descriptors selected for judgment can, at least in principle, be an arbitrary subset of the original ones. However, in this regard it is worth noting that North (2000a: 268–70) divided the CEFR research scale descriptors into three classes: (a) normal descriptors; (b) those showing some variability (differential item functioning) and (c) those showing very good model fit and no variation by context (target language, language of use, educational sector, language region). This subset was dubbed "excellent items" with the suggestion that they might be suitable as anchor items for future projects. These "excellent items" mostly describe aspects of communicative fluency. It was this subset that Jones (2002) used in the ALTE "Can Do" Project, reported in Section 3.3.1. above.

If CEFR descriptors and test items from the examination under study for the same candidates are both included in the data collection, then the difficulty values estimated for the test items can be linked through the difficulty values for the descriptors to the same – CEFR-based – scale.

The new scale could be anchored to the scale from the original CEFR research study in several ways.

- Through descriptor difficulty values:

- by anchoring descriptor-items to their difficulty values in the original study (North 2000a: 358–415);

- by running an independent, unanchored IRT analysis but then equating the new scale and the original CEFR research scale from the relative placement of the CEFR descriptors on the two scales.

- Through cut-offs:

  - by directly anchoring the scale steps to the cut-off points from the original study (North 2000a: 274), shown in Table 1.

Dávid (2007) gives an example of exploiting both the very stable subset of CEFR descriptors ("excellent items") and the CEFR research scale cut-offs. In this project, teacher assessments in relation to the set of CEFR descriptors were used in order to help identify cut-offs for B1, B2 and C1 on the logit scale reported by the local test. In effect, Dávid was able to use the location of the CEFR descriptors on his new logit scale to translate the logit cut-offs on the CEFR research scale into logits on the scale from the local analysis of test items and descriptors. Scale steps were then anchored to these new, local CEFR cut-offs – creating a test scale that reported onto CEFR levels. In this method, it is not the items or the tasks that were linked to the CEFR levels – but the cut-offs themselves.

In this context we should also point out possible problems of translating one scale to another. Logit values are ***not*** a unit of measurement that can be transferred automatically from one context to another; they are the product of a particular approach to data collection and scale construction. There is no reason to expect that a logit scale produced by anchoring together sets of data from teacher assessments in relation to a series of checklists of descriptors using a 0–4 partial credit scale (the CEFR scale research project) would have the same scale properties (e.g. proportional distance between levels, overall length of scale) as a scale produced by an IRT analysis of a suite of vertically linked objectively marked listening tests. The relationship may not be linear. On the CEFR research scale, for example, as shown in Table 1, C1 and C2 are each half as wide as the other levels, when their "plus levels" are included. This phenomenon was also observed in the calibration of the ALTE "Can Do" scale, and may be connected with the difficulty of writing "Can Do" descriptors at the C levels. The same might not be observed with a scale produced from data from objectively scored tests of the receptive skills. The Cambridge ESOL common scale, constructed empirically from objective response data, in fact shows levels decreasing steadily in size as they ascend. This reflects the fact that learning gains decrease proportionally over time: the longer you learn, the less difference it makes.

Table 1: Logit Cut-off Scores of CEFR Levels and "Plus" Levels

Levels		Cut-off	Range on logit scale
C2		3.90	
C1		2.80	1.10
	B2+	1.74	1.06
B2		0.72	1.02
	B1+	-0.26	0.98
B1		-1.23	0.97
	A2+	-2.21	0.98
A2		-3.23	1.02
A1		-4.29	1.06
	Tourist	-5.39	1.10

4.1. Benchmarking with FACETS

One further way of exploiting the descriptors and the scale cut-offs from the CEFR research project concerns the benchmarking of performance samples rather than standard setting for objectively scored test items.

As previously mentioned, an analysis with the program FACETS (Linacre 2008) provides the possibility of taking into account variation in severity amongst judges in order to give a fair rating for a candidate (Linacre 1989), complete with fit statistics and standard errors. The data for the FACETS analysis will probably come from a rating seminar in which a number of experts provide global ratings and possibly analytic ratings on a range of criteria. Manual Section 5.7. recommends the use of Manual Table C2 (CEFR Table 3) or Manual Table C4 (for writing) as the criteria grid for assessing the CEFR level of spoken or written samples of performance respectively at such seminars. Part of the purpose of such seminars is to achieve a consensus through discussion of cases. For the purposes of analysis however it is very important to collect independent ratings *before* any discussion. Such ratings give a better view of the performance of raters and of how the criteria are applied. The calibration to CEFR levels through independent judgments with FACETS can then be compared to the consensus reached by the panel after discussion. Samples that have good model fit (= are not confusing), on which a high consensus was reached, and for which the CEFR level from both FACETS analysis and consensus after discussion are very close, can then be selected as good illustrative samples.

If the descriptor grids mentioned above (Manual Tables C2 or C4) are used as rating criteria – or other grids derived from CEFR descriptors – then the thresholds between the bands on the scale (the scale steps) can be anchored to the cut-points on the logit scale from the CEFR descriptor research scale given in Table 1. IRT ability estimates can then be calculated on the same logit scale as the CEFR descriptors. However, in this situation, the caveat about translating one scale to another still holds. Anchoring to the CEFR research scale will make the scale steps look very even; it would be important to verify whether the unanchored scale shows such clear definition. Poor agreement between raters, different approaches to using the plus levels, or simply the particular distribution of samples for rating, may produce a much less well-defined scale. One should study such effects before anchoring them out of existence.

This FACETS approach was used experimentally at the 1997 Fribourg conference at the end of the Swiss CEFR/ELP research project, implemented in the first international CEFR benchmarking seminar (Jones 2005; Lepage and North 2005a, 2005b, and repeated in the series of such events for different languages that have followed, as well as at the Cross-language benchmarking seminar at Sèvres in 2008 (Breton, Jones, Laplannes, Lepage and North, forthcoming).

5. A Ranking Approach to Cross Language Standard Setting

Users of the pilot version of the Manual and of the toolkit of CEFR illustrative materials for different languages have sometimes questioned the comparability of samples provided for particular languages, purportedly at the same level. There are several legitimate reasons why examples that are supposed to be the same level do not appear to be comparable:

- candidates can be the same global level although they have very different detailed profiles;
- candidates may show aspects of different constructs (adults/teenagers; foreign language learner/ immigrant, etc.);
- the decision on level is a deduction about proficiency on the limited evidence provided by the sample; if the tasks being performed are very different, direct comparison between the candidates becomes difficult.

On the other hand, there is the question of whether the organisations providing these samples really do have a common understanding of the CEFR. Are they working within the same frame of reference or to a local interpretation of it? No amount of CEFR familiarisation and standardisation, or estimation of indices of consistency and agreement, will prove that a given group of experts judging the level for a given language are not bringing their own culturally determined interpretation to the task. In fact, why would we expect them not to?

One way of dealing with the issue is to use plurilingual raters to rate two or more languages as discussed in Manual Section 6.10.3. and as put into practice at the Cross-language benchmarking seminar in Sèvres in June 2008; at the seminar itself, plurilingual raters first rated a number of samples for English and French and then split into subgroups, all rating “anchor” samples for English or French, but going on to rate other samples for a third language (German, Spanish and Italian respectively).

The use of candidate-centred approaches with teacher judgments in relation to CEFR descriptors, using these judgments to anchor the languages to the same scale was discussed in Section 3.2. and Section 4.

Another way of addressing the issue is a ranking approach that exploits the truism that all evaluation involves comparison. Behind every standard there is a norm and setting a standard involves comparing a performance or an item to an internalised norm. The approach put forward in this Manual stresses comparisons which are mediated by the CEFR illustrative descriptors. But every attempt to link a language performance or a test to the CEFR is still an implicit comparison with other language performances and other tests.

The paired comparison method (Thurstone 1927) 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. The problem with the approach is the repetition and sheer number of paired judgments required. Bramley (2005) offers a ranking approach, where more than two objects are compared, as thus an attractive practical alternative. If a set of judges rank a set of 10 samples, so that each sample gets a "score" of 1 to 10 reflecting their ranking, these data can be used not simply to agree a correct rank order, but also to find the relative location of each subject on a Rasch model measurement scale.

A first application of ranking to cross-language comparison took place in the context of the Cross-language benchmarking seminar held at Sèvres in June 2008. Prior to the conference ranking data were collected from the participants on a larger number of samples than could be studied at the actual seminar, 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.

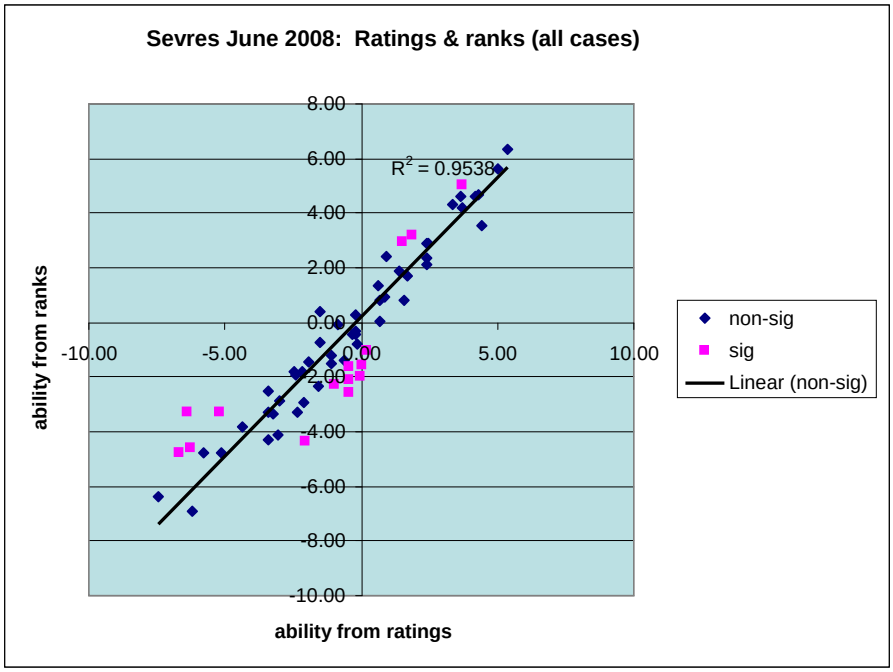

Figure 4: Ranking and Rating Compared (Sèvres June 2008)

Figure 4 compares the abilities estimated from rankings (collected on the website before the seminar) and ratings (at the seminar) for the set of samples submitted to both procedures. The lighter squares are outcomes which are significantly more discrepant than measurement error would allow. Clearly there are some significant differences in the outcomes, but given the fact that the ranking exercise took place before the conference, and was done individually on-line without guidance, discussion or familiarisation with the

procedure – plus an element on confusion in the instructions – this is not surprising. The correlation between the two sets of data is still high at 0.94 for all cases. This first study thus makes the ranking approach look very promising.

Ranking is potentially an extremely useful technique because outcomes are not dependent on the understanding of the CEFR levels, though clearly it remains critical that judges should understand the basis and criteria on which they are to compare performances. It opens up the possibility of constructing a link to the CEFR without a standard setting seminar.

It also suggests a way of expanding the set of "illustrative samples" from the more commonly taught languages investigated at the Sèvres seminar (English, French, German, Spanish, Italian) to lesser taught languages, whilst preserving cross-language comparability. Let us assume that as a result of such multilingual rating conferences we arrive at an authoritative set of samples for widely learned languages. One or more of these languages can then be used as the benchmarks for a ranking exercise vis-à-vis a different language. The cut-offs for the benchmarks can be applied directly to this language and must have the same interpretation. A standard error for such a linking can be estimated (though this is work in progress), enabling us to state with confidence how accurate the outcome is – something which is not possible with standard setting focused on a single language. Cross-language comparison thus is not simply an area for bodies with an explicit interest in multilingual testing. It offers a promising approach of general applicability to any linking exercise.

As with all other benchmarking and standard setting approaches we feel on firmer ground with the performance skills, where judgment is applied to samples of observable behaviour rather than to test items. There remains the issue of whether a comparative approach can be made to work for this case too. 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 devised in which judges would rank items 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. This is an area waiting to be explored.

6. Conclusion

By focusing on the use of scaling techniques and item banking this document has given more emphasis to the location of standard setting and standard maintaining in an operational testing cycle. Having recourse to stable measurement scales facilitates the setting and fine adjustment of standards, and enables us to achieve continuity and consistency in applying them over time. It is certainly not a short cut; a data-based candidate-centred approach requires training large numbers of teachers rather than the members of a small panel. Nevertheless, it does open the prospect of targeting effort most productively. The fundamental challenge for language testers is to write good items. Skilled item writers take years to train, but the effort put into developing that skill and judgment is effort well spent. So is effort put into implementing an operational item banking approach, which will greatly lessen the need for applying judgment in carrying a standard forward over time. Such an approach, however, pre-supposes effective Familiarisation, Specification and Standardisation training of the development team as outlined in Chapters 3, 4 and 5 of the Manual.

In this document we have also tried to show ways in which an external criterion can be integrated into all stages of a linking project. The link to the external criterion is the crux of a linking project, so the more that it can be integrated into the project, the greater the chances of an effective outcome.

Users of this document may wish to consider:

- *how they might exploit the CEFR illustrative items for the skill(s) in question during pretesting*
- *whether they are more concerned with linking one particular standard reported by a test pass (suggests a panel approach) or whether they are concerned to link a range of standards for a series of tests at different proficiency levels (suggests consideration of a scalar approach)*
- *how they address the issue of cross-language standard setting*
- *whether to include teachers in their project, whatever standard setting approach is taken*
- *whether they have the required expertise to conduct a data-based, candidate-centred IRT standard setting study, or whether they could obtain such expertise externally*
- *whether the test(s) under study have demonstrated a sufficiently clear content relationship to the CEFR, through completion of the specification procedures in Manual Chapter 4, for standard setting through such a study to have acceptable validity*
- *whether there is a pool of teachers available with a sufficient awareness of the CEFR as a starting point for it to be feasible to give them the detailed training (Familiarisation; Standardisation Training) necessary for them to give valid CEFR assessments that could be incorporated in the study*
- *whether checklists from European Language Portfolios in the context concerned could be used in such a study (i.e. is the ELP in question a validated one; can the ELP descriptors all really be related back to the original CEFR descriptors or should the bank on www.coe.int/portfolio be consulted for alternatives?*
- *whether other assessment grids could be adapted from (e.g.) CEFR Table 3, Manual Tables C2–4 by selecting relevant levels and categories*

References

Ashton, K. (2008): *Comparing proficiency levels in an assessment context: the construct of reading for secondary school learners of German, Japanese and Urdu. Cambridge Esol Manuscript.*

Baker, R. (1997): *Classical Test Theory and Item Response Theory in Test Analysis. Extracts from: An Investigation of the Rasch Model in Its Application to Foreign Language Proficiency Testing*. Language Testing Update Special Report No 2.

Boldt, R. F., Larsen-Freeman, D., Reed, M. S. and Courtney, R. G. (1992): *Distributions of ACTFL Ratings by TOEFL Score Ranges*. Research Report RR-92-59. Princeton, New Jersey: Educational Testing Services.

Bramley, T. (2005): A Rank-Ordering Method for Equating Tests by Expert Judgement. *Journal of Applied Measurement*, 6 (2), 202–223.

Breton, Jones, Laplannes, Lepage and North, (forthcoming): *Séminaire interlangues / Cross language benchmarking seminar, CIEP Sèvres, 23-25 June 2008: Report.* Strasbourg: Council of Europe.

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

Council of Europe (2007): "*The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities.*" Intergovernmental Language Policy Forum, Strasbourg, 6-8 February 2007, Report.

Dávid, G. (2007): *Building a Case for Euro Examinations: a case study*. Paper given at the Seminar for a joint reflection on the use of the preliminary pilot version of the Manual for "Relating Language Examinations to the CEFR" 2004–2007: Insights from Case Studies, Pilots and other projects. Cambridge, United Kingdom, 6–7 December 2007.

Grin, F. (1999): *Compétences linguistiques en Suisse: Bénéfices privés, bénéfices sociaux et dépenses, rapport de valorisation*. Berne/Aarau, PNR33/CSRE.

Grin, F. (2000): *Fremdsprachenkompetenzen in der Schweiz: Privater Nutzen, gesellschaftlicher Nutzen und Kosten, Umsetzungsbericht*. Bern/Aarau, NFP33/SKBF.

Jones, N. (2002): Relating the ALTE Framework to the Common European Framework of Reference. In Alderson, J.C. (ed.) (2002): *Case studies in the use of the Common European Framework*. Strasbourg: Council of Europe, ISBN 92-871-4983-6: 167-183.

Jones, N. (2005): Seminar to calibrate examples of spoken performance, CIEP Sèvres, 02–04.12.2004. Report on analysis of rating data. Final version. March 1st 2005. Cambridge ESOL internal report.

Jones, N., Ashton, K. and Walker, T. (2007): *Asset Languages: A case study piloting the Manual*. Paper given at the Seminar for a joint reflection on the use of the preliminary pilot version of the Manual for "Relating Language Examinations to the CEFR" 2004–2007: Insights from Case Studies, Pilots and other projects. Cambridge, United Kingdom, 6–7 December 2007.

Lepage S. and North, B. (2005a): *Séminaire pour le calibrage des productions orales par rapport aux échelles du Cadre européen commun de référence pour les langues, CIEP, Sèvres, 2–4 décembre 2004* : Rapport. Strasbourg: Council of Europe DGIV/EDU/LANG (2005)1.

Lepage, S. and North, B. (2005b): *Guide for the organisation of a seminar to calibrate examples of spoken performance in line with the scales of the Common European Framework of Reference for Languages*. Strasbourg: Council of Europe DGIV/EDU/LANG (2005) 4.

Linacre, J. M. (1989): *Multi-faceted Measurement*. Chicago: MESA Press.

Linacre, J. M. (2008): *A User's Guide to FACETS. Rasch Model Computer Program*. ISBN 0-941938-03-4. www.winsteps.com.

North, B. (2000a): *The Development of a Common Framework Scale of Language Proficiency*. New York: Peter Lang.

North, B. (2000b): Linking Language Assessments: an example in a low-stakes context. *System* 28: 555–577.

Szabo, G. (2007): *Potential Problems concerning the Empirical Validation of Linking Examinations to the CEFR*. Paper given at the Seminar for a joint reflection on the use of the preliminary pilot version of the Manual for "Relating Language Examinations to the CEFR" 2004–2007: Insights from Case Studies, Pilots and other projects. Cambridge, United Kingdom, 6–7 December 2007.

Thurstone, L. (1927): A Law of Comparative Judgement. *Psychological Review*, 3, 273–286.

Wilson, K. M. (1989): *Enhancing the Interpretation of a Norm-referenced Second Language Test through Criterion Referencing: A research assessment of experience in the TOEIC testing context.* TOEIC Research Report No 1. RR-89-39. Princeton, New Jersey: Educational Testing Services.

Wilson, K. M. (1999): *Validating a Test designed to assess ESL Proficiency at Lower Developmental Levels*. Research Report RR-99-23. Princeton, New Jersey: Educational Testing Services.

Wilson, K. M. (2001): *Overestimation of LPI Ratings for Native Korean Speakers in the TOEIC Testing Context: Search for explanation*. Research Report RR-01-15. Princeton, New Jersey: Educational Testing Services.

Wilson, K. M. and Lindsey, R. (1999): *Validity of Global Self-ratings of ESL Speaking Proficiency based on an FSI/ILR-referenced Scale*. Research Report RR-99-13. Princeton, New Jersey: Educational Testing Services.