



October 2009

Reference Supplement

to the

Manual for

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

Section I: Cito variation on the bookmark method

Section I

Cito variation on the bookmark method

Frank van der Schoot¹

National Institute for Educational Measurement (Cito)
Arnhem, The Netherlands

Section 6.9 of the Manual for Relating Examinations to the Common European Framework of Reference for Languages (CEFR) describes the Cito variation of the bookmark method. This method uses a rather simple display on which difficulty and discrimination values of all items are presented graphically in relation to the ability scale. An important feature of this display is that panelists are fully informed about the level of mastery for all items in the item pool or test at every point of the ability scale. This informs panelists about the relative difficulty of the items in the test or item pool. Furthermore it prevents panelists making inconsistent decisions. Usually, however, panelists are not familiar with the psychometric concepts involved. Therefore, the standard setting method should be introduced carefully.

In this section we first describe the construction of the display that will be used by the panelists during the standard setting process. Next we will describe how to introduce this display to the panelists and how it can be used in the standard setting process. In the third section we give attention to the standard setting process within the context of the CEFR. The fourth section deals with some practical considerations. Technical aspects of the procedure are discussed in detail in the appendix.

Note that with this procedure of standard setting, the standard is determined on the ability scale directly and not on the scale of the observed scores. To decide whether a particular student has reached the standard or not, two equivalent possibilities are available. In the first method the observed test score is transformed into an estimate of the latent ability and this estimate is compared to the standard. In the second method the standard set on the latent scale is transformed into a standard of the observed test scores, and a particular test score is compared to this standard. Both methods are discussed in the appendix.

I.1. The construction of the item display

We assume that an IRT-calibrated set of items is available which is validated for the proposed ability. The range of the ability scale for which the item pool is informative encompasses the presumable standard score for the CEFR decision level. Although in all Cito applications of this standard setting method the IRT model used has been OPLM (see Section G of the Reference Supplement), the method is equally well applicable with other IRT models, for example, the three parameter logistic model.

Next we adopt a general decision rule for the level of mastery of single items. For example, as is also suggested in section 6.9, ‘borderline mastery’ of an item is designated if the probability to get an item correct is 50%, i.e. response probability (RP) = 0.50. Students with an ability level which corresponds to RP less than 0.50 have an insufficient mastery of an item. Full

¹ The present chapter has been included in the ‘Reference Supplement’ with the kind permission of the author. Copyright remains with the author. Correspondence concerning this chapter or the reproduction or translation of all or part of it should be sent to the author at the following address: Frank van der Schoot, Christinastraat 43 6862 GK Oosterbeek The Netherlands, frank.vanderschoot@planet.nl.

mastery of an item is reached at ability levels which correspond to RP of 0.80 and higher. Ability levels that correspond to RP between 0.50 and 0.80 are considered as moderate levels of mastery. This segment of the ability scale is briefly indicated as the P50-P80 segment. Thus for each item, the ability scale is divided into three segments: (1) a segment that corresponds to insufficient or poor mastery where $RP < 0.50$, (2) a segment that corresponds to moderate mastery where RP falls between 0.50 and 0.80 and (3) a segment that corresponds to full mastery, where RP is greater than 0.80. Choosing an 80% probability for a correct answer as the definition of full mastery is essentially an arbitrary decision, and there are no psychometric reasons for doing so. If one considers that this criterion is too harsh, one might lower it, for example, to 75%. As long as this choice is well documented and clearly communicated to the panel members, there is no problem with it.

The display that is used in the standard setting process shows the P50-P80 segments of all items, ranked on the difficulty parameter, where RP equals 0.50. (see Figure 6.4 in the Manual or Figure 5 in this section). In the appendix, it is shown in detail how the end points of these segments can be computed. There it will also appear that the point with $RP = 0.50$ does not coincide with the difficulty parameter when the three parameter logistic model is used. When using P50-P80 segments it is recommended – especially when these segments are rather long caused by relative low discrimination parameters - to mark the position of RP 0.65 in these segments. This divides the P50-P80 segment into a segment with rather weak mastery to the left and a segment with rather good mastery to the right (see Figure 6 in this section). Notice that the RP65 point is not the midpoint of the RP50 and RP80 points, but in general it will be very close to it. For practical applications it is not important whether the RP65 point or the midpoint is displayed.

It is further recommended – see also section 6.9 of the Manual - to use a convenient scale having no negative values, and an easy-to-understand unit, avoiding interpretations in terms of percentages, and being fine-grained enough to require provisional standards expressed as whole numbers. In the examples below we will use a transformation of the latent scale to an ability scale that ranges from 100 to 400. A detailed explanation on how such a transformation is computed is given in the appendix.

1.2. Introduction of the displays to the panel members

As was noted earlier, panelists are usually not familiar with the basic psychometric concepts of item response theory involved in constructing the display. It is of course most important that they fully understand the display that will be used in the standard setting process as their response sheet.

First we have to explain the concept of response probability (RP) and its relation to an ability scale: RP varies between 0 and 1 and increases with ability level. Figure 1 shows a typical item response function. At every point on the ability scale we can read the probability of a correct response for this item along the y-axis. For example, students with an ability score of 250 have a chance of less than 20% to answer this item correctly, while students with an ability score of 300 have a chance of 80% to answer this item correctly. Panel members are told that for all the items they will have to consider, such a curve has been computed by the application of a psychometric theory on the data collected for the calibration study. It is important that panel members are convinced that these curves are based on empirical data, and are not mere theoretical guesses.

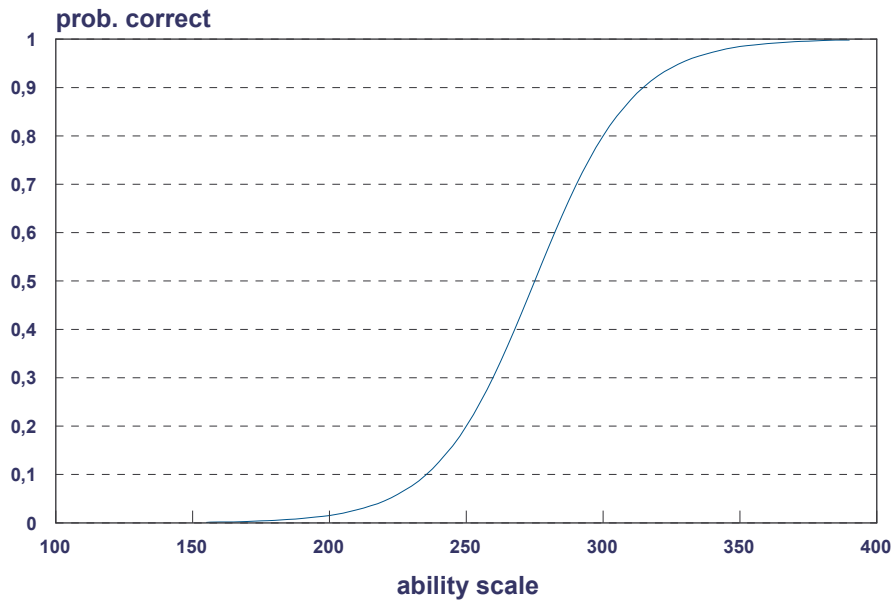


Figure 1. An example of an item response function

We choose two critical points on this scale, corresponding to the P50-P80 segment that we will use in the display, i.e. the first point at $RP = 0.50$ and a second point at $RP = 0.80$. This divides the ability scale for this item into three segments as is shown in Figure 2. These segments represent the three levels of item mastery:

- poor or insufficient mastery when RP is less than 0.50: a correct answer is obtained in less than 5 out of 10 cases;
- full mastery when RP is 0.80 or higher: a correct answer is obtained in 8 or more out of 10 cases;
- and moderate mastery when RP lies between 0.50 and 0.80.

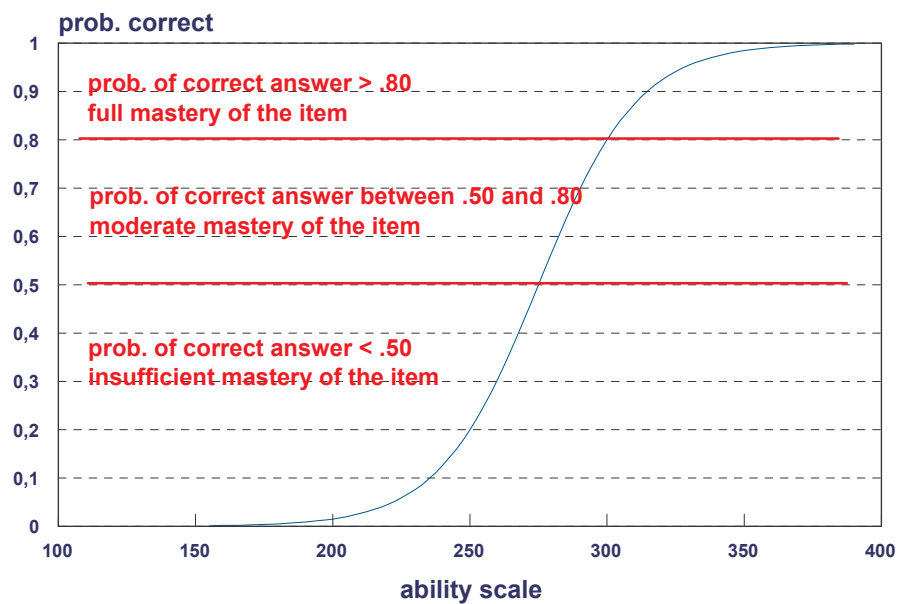


Figure 2. Three levels of mastery in relation to the item response function

Figure 3 shows also the projection of RP's 0.50 and 0.80 on the ability scale. RP equals 0.50 at ability score 275, and RP equals 0.80 at ability score 300. Thus, students with an ability score less than 275 show insufficient or poor performance on this item, students with an ability score between 275 and 300 show moderate performance and students with an ability score higher than 300 show full mastery of this item.

The range of moderate performance is specifically indicated by the horizontal line or block, identified as item A. Thus, the ability levels to the left of segment A correspond to poor performance of item A, ability levels to the right of this segment correspond to good performance of item A, while the segment itself corresponds to moderate performance on this item. If a panelist judges this item too difficult for the standard, than the standard score must be lower than 275. On the other hand, if he/she judges that the standard implies full mastery of this item, the standard will be higher than 300.

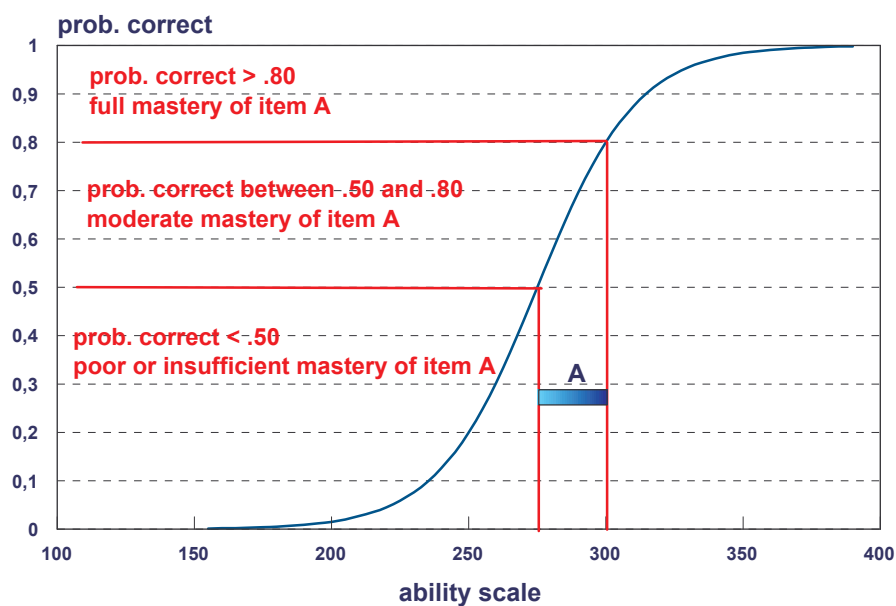


Figure 3. P50-P80 segment for item A

Figure 4 shows the response function of a second item, item B, on the same ability scale. The range of the P50-P80 segment for item B lies between ability scores 200 and 250. The ability scale is now divided into five segments:

- students with an ability score less than 200 show poor performance on both items;
- students with an ability score between 200 and 250 show moderate performance on item B but poor performance on item A;
- ability scores between 250 and 275 correspond to full mastery on item B but still poor performance on item A;
- ability scores between 275 and 300 reflect full mastery on item B and moderate mastery on item A; and
- students with ability scores higher than 300 show full mastery on both items.

Now suppose we have to indicate a standard, e.g. for the A2/B1 decision, on the basis of this two-item test and B1 implies full mastery of item B but that it is not necessary to master item A even to a moderate extent. Then, of course, our decision for this standard lies between

ability scores 250 and 275. Additional items in the display will help us to further define the position of the standard .

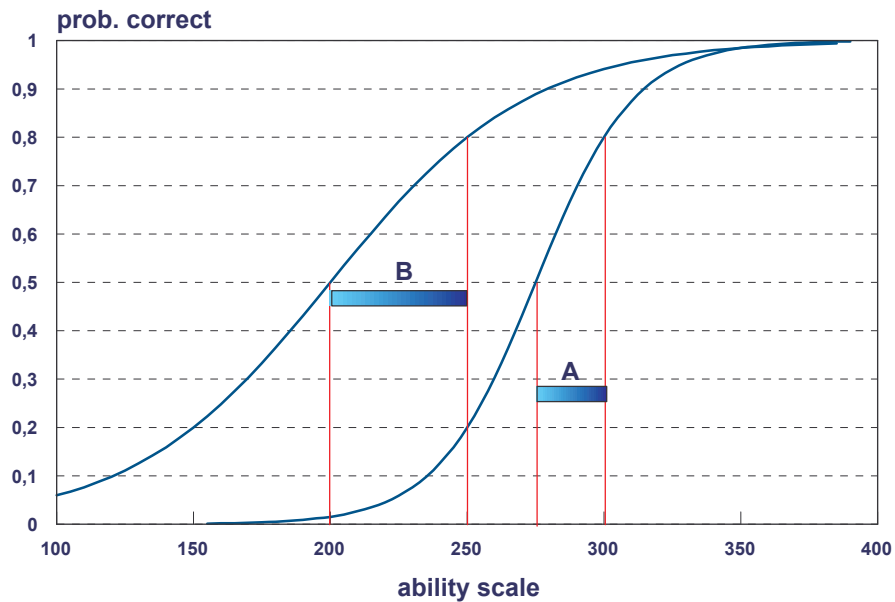


Figure 4. P50-P80 segment for an additional item B

After the presentation of Figure 4, it may be useful to point out how difficulty and discriminatory power of items are displayed in the graph. The horizontal segment representing item A is situated more to the right than that for item B, and this difference reflects differences in difficulty: the more to the right the segment, the more difficult the item is. Differences in discrimination are reflected in the length of the line segment: the longer it is, the less the item discriminates. In the example item B discriminates less well than item A, which can also be seen from the item response curves: the curve for item A is steeper than the curve for B.

If the Rasch model is used, all line segments will have the same length. With the two parameter model or OPLM, segments of different lengths will appear and the length depends only on the discrimination parameter, while in the three parameter model the length of the line segments depends in a quite complicated way on the guessing and the discrimination parameters. See the appendix.

Figure 5 shows a display of a short 8-item test. Panelists are now asked to answer some questions individually or in small groups. This is an important exercise to learn to interpret the display and gives important feedback about the effect of the introduction so far. Of course the correct answers are discussed extensively with the panelists.

Examples of questions which the panelists should be able to answer now, are:

- Which item in this test is the easiest one? Which item is the most difficult one? Give arguments in support of your answers.
- Which students have insufficient mastery of item 3? Which students show full mastery on this item?
- What is the response probability for students with an ability score of 250 on item 5? What does this mean? To what extent do these students master item 5 ?

- Which items of this test are mastered fully by students with an ability score of 250? Which items are insufficiently mastered by these students? Answer these same questions also for students with an ability score of 200 and an ability score of 350.
- Suppose that a standard level has been set at the ability score of 230. Which items of this test should be mastered fully or moderately? Which items are judged to be too difficult for this standard?

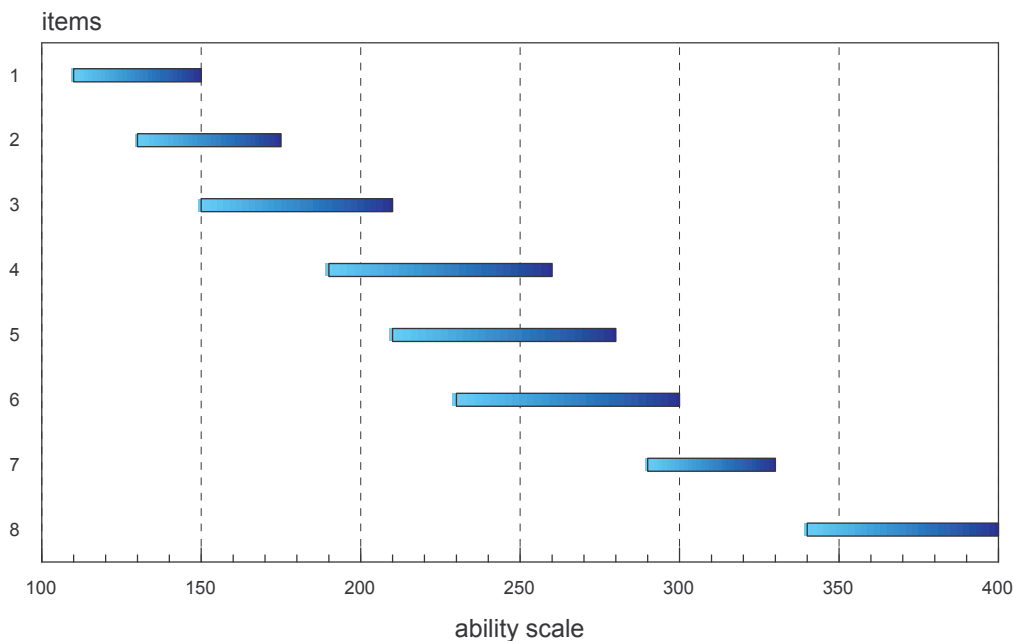


Figure 5. Item map for an 8-item test

We continue the introduction until it appears that all panelists fully comprehend the display. Then we can start to explain in more detail the decision process to define the standard itself.

The aim of the standard setting process is to answer the following question:

To what extent should the items of this test or item pool be mastered at the decision level under study, e.g. A2/B1-decision level. First you make a decision on each separate item. After that you should decide on an ability score that best represents the decisions on all individual items.

Figure 6 shows the display with decisions on individual items by an individual panelist. This panelist has studied the items and has decided for each item to what extent it should be mastered. The cross marks in this display show his/her decisions. Note also that in this case we also use indicators of $RP = .65$, dividing the section of moderate performance into weak and rather good performance.

This panelist has decided that the standard implies full mastery of the first three items. We disregard the specific position of these three marks on the ability scale. For all three items, the decision is the same: they should be mastered fully. On the other end of the scale, the cross mark indicates that item 8 is judged to be too difficult in relation to the standard. According to the marks for items 4 to 7, the standard score lies within the range 250 to 300. If, however,

moderate performance on item 7 is required, then this panelist also has to decide for full mastery on items 4 and 5, and even on item 6. On the other hand, if full mastery of these items is not required, then our panelist has to decide that there is insufficient performance on item 7.

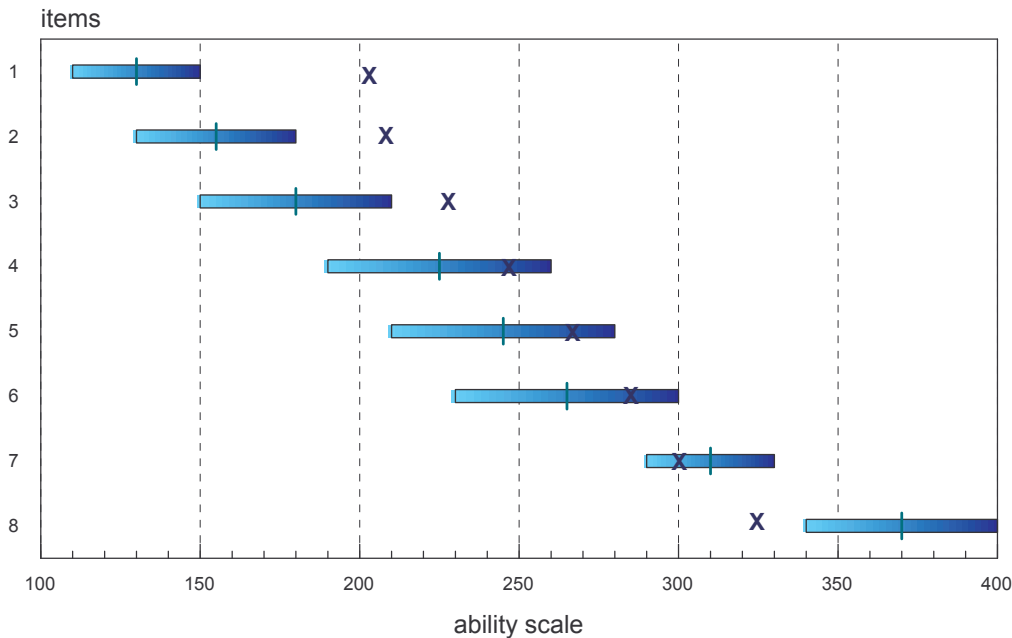


Figure 6. Item map, indicating RP65 and judgments on each item

Of course this is a difficult decision process in which panelists have to weigh the importance of mastery of a subset of items with respect to the standard to be set. Sometimes conflicts can arise, e.g. when a panelist decides for full mastery of item 6 and poor mastery of item 7. But finally he/she has to indicate one specific point as his/her best decision for the ability level for the standard to be set.

A try-out

After this introduction it is important to train the panelists in using the display. We now provide them with a set of items and its corresponding display. This can be a short test, e.g. 10 or 15 items, preferably with items that are not used in the standard setting procedure itself. Sometimes it is possible to use items from a different item pool, which is not directly related to the subject under study. The focus of this try out should be on the decision process and the proper use and interpretation of the display.

From this try out panelists will learn that they have set different values on the scale as the standard. Therefore it is recommended that the standard setting procedure has two or three stages. The standard setting procedure can start when all participants fully comprehend the display and feel sure in using it.

I.3. The standard setting procedure

Stage 1

Now we can start with the standard setting process itself. All panelists receive a booklet with the items and the corresponding display. The items are ranked on the difficulty parameter, in

the booklet as well as on the display². Sometimes it appears difficult to rank the items in the booklet in this way, for example, in the case of a reading test. Typically, such a test contains several short texts and each text is followed by several items of which some appear to be rather easy and others to be rather difficult. In this case we rank the items for each text in the order of difficulty but the number of each item corresponds to its position in the total item pool. Thus, for example, the first text in the booklet can be followed by four items, numbered 2, 5, 10 and 24, indicating their relative position in the total item pool.

In the first stage, each panelist studies thoroughly the items and decides individually for the position of the standard on the ability scale. It is expected that panelists will indicate different positions.

Stage 2

In the second stage of the procedure, panelists are put together in groups to discuss the decisions each panelist has made. The intensity and effectiveness of these discussions are enhanced by setting up small groups of about 5 to 6 panelists. They interact and discuss the arguments for their decisions.

In the end, each panelist indicates a second score for the standard, taking into account the result of the discussion. Generally the range of these scores will be significantly reduced. Of course one can also ask each group to reach a group decision for the proposed score.

At this stage, the panel members receive normative information (see Section 6.2.1. of the manual), where the most important purpose is to reflect on one's own decisions and to reach a better understanding of the considerations that enter into quite a complex decision making process. It is hoped that this reflection will be facilitated through discussing differences in the decisions reached thus far. Therefore, it is important for this stage to consider very carefully the way the discussion groups are formed. If a discussion group is formed where all members agree to a large degree in their first decision, not much thinking or useful discussions can be expected. For the team leader or coordinator, therefore, it is important to look carefully at the first decision and to form groups such that there is sufficient disagreement to provoke serious discussions.

After gathering the individual or group decisions, the standard can now be determined, e.g. the median or the mean of the individual standards (see Section 6.3.4. of the Manual).

Stage 3

However, by adding a third stage to the standard setting procedure we can present the individual or group decisions to all participants. For example, we now can present a display with the interquartile range of the decisions superimposed (see Figure 7). Further discussions with all participants can result in a standard on which all individuals or groups agree. Sometimes the results of a particular reference population are available. In that case we can relate the standard to the performance distribution of this population. Figure 8, for example, displays quite a lot of information. Here is some explanation about this figure.

² If the three parameter model is used, it is advisable to rank order the items on their P50 value.

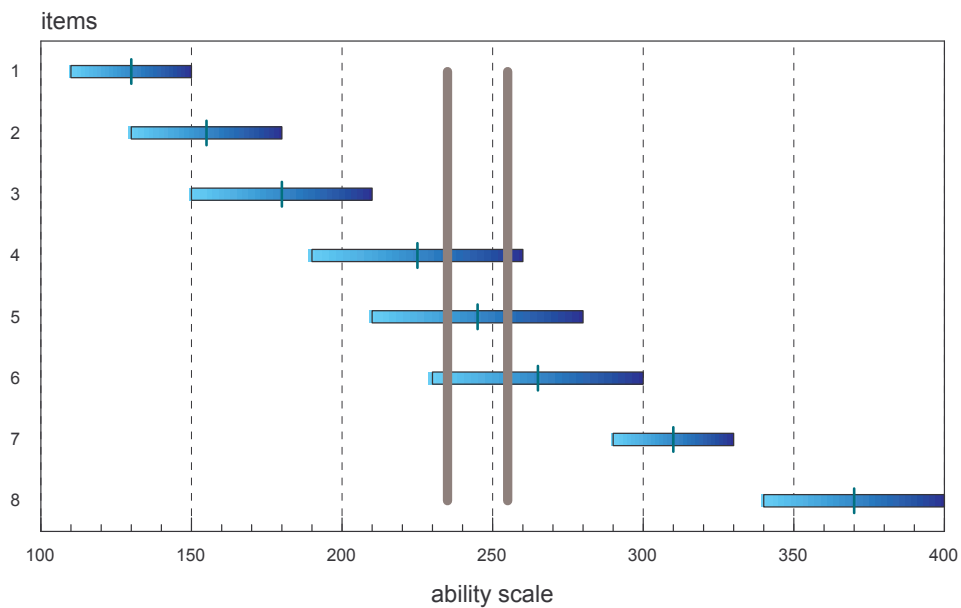


Figure 7. Item map with interquartile range of judgments superimposed

- On top of the figure the percentiles 10, 25, 50 (the median), 75 and 90 of some reference population are indicated. The shorter vertical lines help in judging the numerical value of these percentiles. The 25th percentile, for example, is approximately 216, and the median is 250.
- The item segments (horizontal bars) can be related to the percentiles. For example, the P80 point of item number 1, with value 150, is a lot smaller than the 10th percentile (about 185), meaning that more than 90% of the population fully masters this item. For item number 7 we can deduce that more than 75% percent of the population has insufficient mastery of this item.

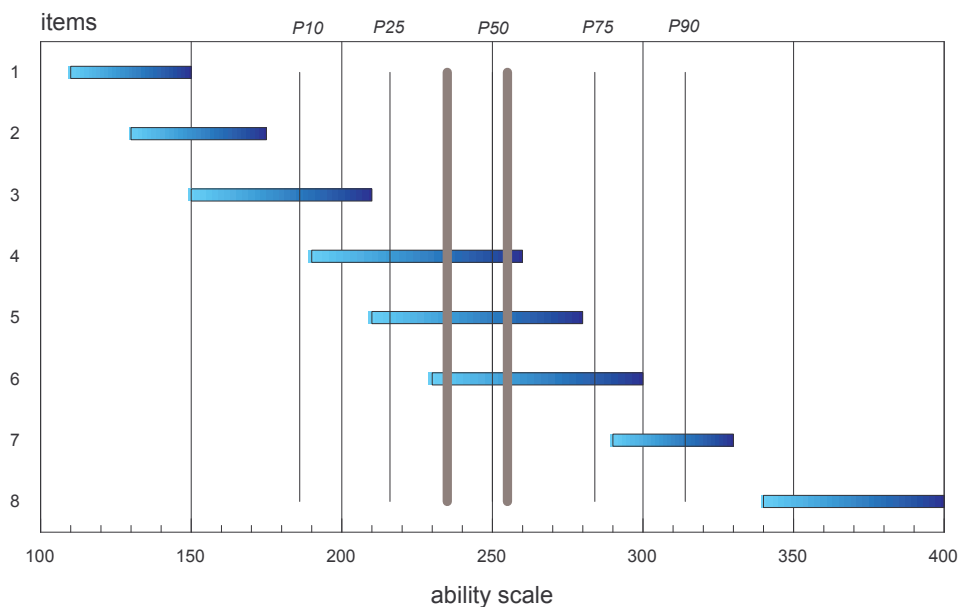


Figure 8. Item map with interquartile range of judgments and with five percentile points of the ability distribution for a population of reference

- The two thick vertical lines (with horizontal values of about 235 and 255, respectively) display the interquartile range of the final decisions after the second round. This means that 50% of the panel members arrived at a standard between these two values, 25% had a standard lower than 235 and 25% came up with a standard higher than 255. This range gives a picture of the remaining disagreement between panel members after a thorough discussion round, and this disagreement certainly must appear in the report on the standard setting procedure.
- The most important feature of Figure 8, however, is that it provides impact information (see Section 6.2.1 of the Manual). If the median of the individual decisions is taken as the final group decision, it is seen to be very close to 250, which is also the median of the ability distribution in the population. If this standard were to be used for deciding on success or failure in an examination, it follows that about 50% of the population would fail the examination. This is important information for the panel members, who might wish to revise their decisions, but also for the authority that is responsible for the final decision, and who might change the final advice from the panel because of reasons outside the competency of the panel members

I.4 Practical considerations

It may happen that the test or examination for which one wishes to set one or more standards contains so many items that the basic display used by the panel members is jam-packed with line segments so as to become hard to read and confusing. However, if the test is well calibrated using an item response model, there is no need to display the segments of all items, a well chosen subset will suffice. By selecting the subset one should be careful in two respects:

- There must be sufficient variation in difficulty so that it is likely that the standard(s) to be set fall well within the range of the line segments presented to the panel members;
- The subset chosen must be representative for all the items in terms of content and relevant Can-do statements from the CEFR.

By using subsets of items, one can even set standards that are applicable for a whole (calibrated) item bank, and in principle the standards – cut-off scores on the latent ability underlying the performances – can be applied to any test drawn from the item bank. The way to do this is described in detail in the appendix.

The use of subsets of items also offers an excellent opportunity to collect evidence for the validity of the standard setting procedure (see Chapter 7 of the Manual). After the second stage, one might give another display, using a moderate number of items (10 to 12, say) which were not used in the first display, and ask the panel members to set the standard(s) for this short test too. To eliminate memory effects, it is advisable to use another scale for this second test. For example, if in the first display a scale is used running from 100 to 400, one could now use a scale running from 350 to 650. The two scales can easily be converted to one another and in an extra discussion round we can show for each panel member the standards set using either subset of items on a single common scale. Substantial differences for a single panel member between the two standards set shows a lack of intra-judge consistency.

A final question concerns the issue of setting multiple standards using the same test or item pool. As an example, suppose that a standard has to be set for A1/A2 and for A2/B1 using the

same test. Then one could proceed in essentially two ways: either both standards are set in one session using the same display or two sessions are organized, one for each standard and the display presented to the panel members can, and probably will, differ.

The advantage of combining the two standards in one session and using the same display is that the panel members immediately see the consequences of their decisions. It might turn out, for example, that a panel member sets the two (provisional) standards so close to each other that he/she is immediately confronted with the fact that there is barely a distinction between the two standards, and that therefore at least one, and presumably both, standards have to be reconsidered.

But there is also a disadvantage associated with the combined procedure, which may be seen clearly when one considers Figure 6. Suppose this figure is the display for a single standard, and suppose furthermore that all panel members set their provisional standard between 110 and 120, implying that in their view all but one of the items are far too difficult for a student with ability equal to the standard. One would hardly have any confidence in this standard, because the majority of the items to be considered is not very relevant to the decision at stake. Conversely, this means that – as a coordinator of the standard setting procedure – one has to have a fairly good idea of the region (on the ability scale) where the standard will probably be set, and choose the items to be displayed in such a way that a substantial proportion has to be mastered at the standard and another substantial proportion of too difficult items, such that the panel members can come to a well-considered decision.

If one tries to comply to this rule for two or more standards at the same time, and if these standards are quite far apart, then one will have to display many items, and for each standard considered separately, quite a lot of the displayed items will be trivial, and this might confuse the panel members. Therefore it is recommended to organize separate sessions for each standard.

Appendix

This appendix contains three sections. In the first section it is shown how the RP50 and RP80 points are determined in the Rasch model, the two-parameter logistic model (or OPLM) and the three parameter logistic model. In the second section the transformation to a more suitable scale is discussed and in the last section it is shown how a standard, defined in terms of the latent scale, can be expressed as a standard (cut-off score) in the domain of the test scores.

A.1 Finding the points RP50 and RP80

Remember that RP50 is the point on the ability scale (a value of θ ; see appendix G of the Reference Supplement) for which there is a probability of 0.50 for a correct response. Similarly, RP80 is the ability which gives a probability of 0.80 for a correct response.

The Rasch model

In the Rasch model RP50 for item i is the value of θ for which the following equation is true:

$$P(X_i = 1 | \theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} = 0.5 \quad (1)$$

But then, for the same value of θ the probability of an incorrect response must also be 0.50, and we can write a second equation:

$$P(X_i = 0 | \theta) = \frac{1}{1 + \exp(\theta - \beta_i)} = 0.5 \quad (2)$$

Dividing equation (1) by equation (2) gives

$$\text{RP50: } \frac{P(X_i = 1 | \theta)}{P(X_i = 0 | \theta)} = \exp(\theta - \beta_i) = 1 \quad (3)$$

To get rid of the exponential function (exp) in the right-hand equation of (3), we ‘undo’ it by applying its inverse function, which is the logarithm (ln), meaning that for any number x it holds that $\ln[\exp(x)] = x$. So, taking logarithms of both sides of the right-hand equation in (3) gives

$$\theta - \beta_i = \ln(1) = 0$$

from which we find immediately that the solution is $\theta = \beta_i$ or RP50(item i) = β_i .

To find RP80, we can use the very same technique: for equation (1) we fill out 0.80 in the right-hand side of (1) and 0.20 in the right hand-side of (2) and analogous to equation (3) we find

$$\text{RP80: } \frac{P(X_i = 1 | \theta)}{P(X_i = 0 | \theta)} = \exp(\theta - \beta_i) = \frac{0.80}{0.20} = 4 \quad (4)$$

and this gives the solution RP80(item i) = $\beta_i + \ln(4) = \beta_i + 1.386$. If RP75 is used as a criterion for full mastery, then one finds easily that RP75(item i) = $\beta_i + \ln(0.75/0.25) = \beta_i + \ln(3) = \beta_i + 1.099$.

The two-parameter logistic model and OPLM

To find RP50 in this model, we find analogous to equation (1) that

$$P(X_i = 1 | \theta) = \frac{\exp[a_i(\theta - \beta_i)]}{1 + \exp[a_i(\theta - \beta_i)]} = 0.5 \quad (5)$$

where a_i is the discrimination parameter, a positive number. Likewise it must hold that

$$P(X_i = 0 | \theta) = \frac{1}{1 + \exp[a_i(\theta - \beta_i)]} = 0.5 \quad (6)$$

Dividing (5) by (6) gives

$$\text{RP50: } \exp[a_i(\theta - \beta_i)] = 1 \quad (7)$$

Taking logarithms of both sides of (7) gives

$$\text{RP50: } a_i(\theta - \beta_i) = 0 \quad (8)$$

and since a_i is positive, the only solution is RP50(item i) = β_i .

For RP80, we find analogous to (7) that

$$\text{RP80: } \exp[a_i(\theta - \beta_i)] = 4$$

Taking logarithms and solving for θ yields

$$\text{RP80(item } i): \theta = \beta_i + \frac{\ln(4)}{a_i}.$$

The three parameter logistic model

To find RP50 for an item, one has to find the value of θ such that

$$P(X_i = 1 | \theta) = c_i + (1 - c_i) \frac{\exp[a_i(\theta - \beta_i)]}{1 + \exp[a_i(\theta - \beta_i)]} = 0.5 \quad (9)$$

where c_i is the so-called guessing parameter, a number in the interval $[0,1]$. From (9) one can see immediately a possible problem: the middle term consists of a sum of two quantities, neither of which can be negative. Now suppose that $c_i = 0.55$, then the equation says that we have to add to 0.55 a non-negative number such that the sum is 0.5, which is not possible. Therefore, RP50 only exists if $c_i < 0.50$, which we will assume to hold in the sequel.

The solution is found in two steps. First we define

$$\kappa_{50} = \frac{0.5 - c_i}{1 - c_i} \quad (10)$$

and we note that, since $c_i < 0.5$, it holds that $0 < \kappa_{50} < 1$. Then the last equation in (9) can be written as

$$\frac{\exp[a_i(\theta - \beta_i)]}{1 + \exp[a_i(\theta - \beta_i)]} = \kappa_{50} \quad (11)$$

which is similar to the last equation in (5). From (11) it is not difficult to verify that

$$\frac{1}{1 + \exp[a_i(\theta - \beta_i)]} = 1 - \kappa_{50} \quad (12)$$

and dividing (11) by (12) we find that

$$\exp[a_i(\theta - \beta_i)] = \frac{\kappa_{50}}{1 - \kappa_{50}}. \quad (13)$$

Taking logarithms of both sides in (13) and solving for θ gives

$$\text{RP50: } \theta = \beta_i + \frac{1}{a_i} \times \ln \left[\frac{\kappa_{50}}{1 - \kappa_{50}} \right]. \quad (14)$$

To find RP80, we replace κ_{50} in (14) by κ_{80} , and this can be found by replacing 0.5 by 0.8 in (10). Similarly for RP75.

A.2 Transforming the latent scale

Most IRT software reports its results on a scale such that estimated item parameters (difficulties) and estimated θ -values assume negative as well as positive values, usually contained in a small range around zero, such that decimal numbers are required to make necessary distinctions. For panel members not very used at doing mathematics, this may be confusing, and especially negative numbers may be conceived as something bad, that has to be avoided. Therefore, it is advisable to use a scale that has no negative numbers, and where using only integer numbers does not lead to any important loss in accuracy. Moreover, it is to be advised to avoid scales which can be easily confused with percentages. In the preceding examples, a scale has been used running from 100 to 400.

The transformation used is a *linear transformation*, meaning that we use a rule that transforms any value on the original θ -scale in a new, more suitable value, V , say. The rule for a linear transformation states that, for any value θ the corresponding value V is found by

$$V = B \times \theta + A$$

In this section it is explained how to find the values of the coefficients B and A .

Suppose one wants to do a standard setting using 20 items. The RP50 and RP80 points for some items *on the original scale* are displayed in Table A.1. The items are ordered in increasing value of their RP50.

Table A.1 Original and transformed values for RP50 and RP80

item	Original scale		Transformed scale	
	RP50	RP80	RP50	RP80
1	-2.78	-1.56	109.1	159.6
2	-2.23	-1.14	131.9	177.0
...
19	2.95	4.11	346.2	394.2
20	3.07	4.04	351.2	391.3

If we were to display the item segments using the original scale, we would probably use a minimal displayed value which is a bit smaller than the smallest RP50 to avoid that the segments stick to the left vertical axis. Assume that we would choose -3. Similarly, for the largest displayed value we would choose a value a bit larger than the largest RP80, and in the example 4.25 seems a reasonable value. But we want a transformation such that the smallest displayed value is 100, say, and the largest is 400. Then we can write down two equations:

$$B \times (-3) + A = 100 \quad (15)$$

and

$$B \times 4.25 + A = 400 \quad (16)$$

From (15) we find that

$$A = 100 - B \times (-3) \quad (17)$$

Substituting the right-hand side of (17) for A in (16) we find that

$$B \times 4.25 + 100 - B \times (-3) = 400$$

and solving this for B , we find that

$$B = \frac{400 - 100}{4.25 - (-3)} = 41.379$$

Using this result in (17), we find for A :

$$A = 100 - 41.379 \times (-3) = 224.137$$

Having found the value for both coefficients A and B , the linear transformation can be applied to all RP50 and RP80 values on the original scales. The transformed values are filled out in the two rightmost columns of Table A.1.

We generalize this result for arbitrary values. Using lower case letters for the original scale and upper case letters for the transformed scale, the lowest and highest values on the original scale are symbolized by l and h respectively on the original scale and by L and H on the transformed scale. Then the coefficients of the transformation are given by

$$B = \frac{H - L}{h - l} \quad (18)$$

and

$$A = L - B \times l. \quad (19)$$

Notice that the panel will set the standard(s) on the transformed scale, and for further calculations it may be useful to have the standard on the original scale. Denote the standard on the transformed scale by V_c and on the original scale by θ_c . Because the transformation from the θ -scale to the V -scale is linear, the back transformation is also linear, i.e., there are two coefficients a and b such that

$$\theta_c = b \times V_c + a$$

The values of these coefficients are given by

$$b = \frac{h - l}{H - L}$$

and

$$a = l - b \times L.$$

In the example given, this gives $b = 7.25/300 = 0.0242$ and $a = (-3) - 0.0242 \times 100 = -5.42$. Suppose $V_c = 255$, then $\theta_c = 255 \times 0.0242 - 5.42 = 0.751$.

A.3 Decision making

The practical application after standard setting is to make decisions on individual student performances, i.e., deciding whether a student has passed the exam, or can be given the B1 qualification or not. In this section three methods are discussed on how this can be done. It is assumed that the standard has been expressed as a value on the original scale.

The first method consists in estimating the latent ability of the student. Such estimates are given in most IRT software. In Section G of the Reference Supplement, maximum likelihood and weighted maximum likelihood estimates are discussed. Usually, one will take the decision that the student has passed the exam or deserves the qualification B1, etc. if the estimated value of the latent ability of the student is greater than or equal to the standard. Although such a method is good, it has the disadvantage that it may not be transparent to the students, because it is fairly complicated to explain in simple wordings how this estimation of the latent ability works.

Another method is to translate in some sense the latent ability to the scores one can obtain on a test. This will be discussed only for the case that all items are binary.

The simplest, but least accurate method is to express the score as a raw score, i.e. the number of items correct. As is explained in section G, the item response function is the probability of a correct response for a value of θ , but for binary items, it is also the expected score on the item (see Section C). So we can write

$$f_i(\theta) = P(X_i = 1 | \theta) = E(X_i | \theta) \quad (20)$$

This holds for all possible values of θ , and in particular for the value of θ which corresponds to the standard, θ_c , say. So for a person with an ability equal to the standard, his expected

score on item i is $f_i(\theta_c)$. Since the expected (raw) score on a test is the sum of the expected item scores, it holds that at the standard the expected (raw) score is

$$E(S | \theta_c) = \sum_i f_i(\theta_c) \quad (21)$$

To compute this expected score, one has to evaluate (21) using the estimated item parameters from the calibration. This can be done for any IRT model used for the calibration. Usually the expected score will be a decimal number, say 28.23. This means that a person having a θ -value equal to the standard will obtain on average a raw score of 28.23, and to obtain this score or a higher one in a test administration the number of items correct must be at least 29, which is the raw score standard, i.e., the rounding is upwards. But see also the discussion in Section 6.3.4 of the Manual about rounding, where it is argued that downwards rounding might be the better choice in some cases.

The third method is only applicable when the IRT model used is the two-parameter model or OPLM. The decision is based not on the raw score but on the weighted score. The weighted score S_w is the sum of the discrimination indices for the correctly answered items and the expected weighted score is given by

$$E(S_w | \theta_c) = \sum_i a_i f_i(\theta_c)$$

This method gives more accurate results than the previous one, but it may be less transparent to the students. In any case, if it is decided to use it, then the weights for each item should be known by the students, and it is not always easy to explain why some items have a greater weight than others.