Chapter 16

Levels of Performance

Charles Alderson

Introduction

As we have seen in earlier chapters (especially Chapter 8), it is difficult to say what 'levels' the pilot tasks were at, partly since these are not defined by the Ministry – what do 'Basic', 'Intermediate' and 'Advanced' actually mean? – but also because the Hungarian English learning population is highly heterogeneous – see Fekete et al, (1999). This heterogeneity means that it is probably impossible to define only two levels for the Year 12 School-leaving examination (which is what OKI is supposed to deliver) without at the same time discussing pass marks and pass rates.

In addition, it is unclear what is meant by 'level' anyway. If Year 12 students are to achieve an Intermediate level, does that mean that only those students who get a mark of 5 on the traditional Hungarian scale are to be considered at the appropriate level? Or are all who get at least a 2 to be considered at the right level? Clearly this is a crucial issue, especially in light of the fact established by the Baseline Study (Fekete et al, 1999) that virtually everybody gets at least a 3, and over 85% of the population get at least a 4. Indeed, the fact that virtually nobody 'fails' the exam means that it is for all practical purposes worthless.

One possible meaning of the terms Basic, Intermediate and Advanced is that they refer to the levels as defined by the State Foreign Language Examinations (SFLEB). However, if this is the intention, there are a number of consequences, since the Baseline Study showed high failure rates for secondary school students on the State Foreign Language Examination Board examinations, even for those who attempt this exam, and of course, many more do not even make the attempt. Indeed one might argue that it is precisely because students can and do fail the SFLEB exams that these have currency, despite their somewhat idiosyncratic content and methods, and the lack of evidence for their reliability and validity.

The Baseline Study showed clearly that English teachers were ambivalent about the plans to reform the exam. On the one hand they considered the existing exam to be in urgent need of reform, since it is hopelessly outdated in terms of content and method. And importantly, since everybody passes, it has no currency – it is worthless. Teachers would like the exam to have currency. At the same time, however, they like the existing exam because their students pass, and they feel that low pass rates would reflect badly on their teaching! They are well aware of the dilemma that if the exam is to have currency, students will have to fail.

The Government wishes to reform the examination system, yet has not addressed the issue of what standards will apply and whether pass rates can be expected to be considerably lower than at present. In such a policy vacuum one option is simply to sit still and wait for decisions. However, we believe that this is irresponsible and so we have undertaken a study of the levels of performance on some of the pilot tasks to see what implications can be drawn for definitions of levels and maybe even pass rates.

Method

There are two different ways in which we explored levels of achievement on the pilot tasks. We needed to know which tasks were thought to be Basic, which Intermediate, and which Advanced. We also had to estimate the difficulty of the tasks themselves, since a low score on a difficult task means something quite different from a low score on an easy task.

First, we used a judgemental method, which involved recruiting nine experts who were familiar both with the levels of the Council of Europe Common European Framework of reference (CEF) and the levels of Hungarian learners of English, and asking them to estimate levels of difficulty of tasks and items. Such judgements would potentially benefit item-writers as well as enabling us to explore levels of achievement.

Second, we were able to compare empirical pilot task difficulties with those of the CITO anchor items (see Chapter 4), since the level of the CITO anchors is known in comparison to the Council of Europe's CEF – they are said to be at A2 or Waystage. This involved calibrating the pilot tasks with the CITO items, and calculating levels. We were then able both to estimate task difficulty empirically, and to calculate student ability scores, on the same scale. We could then inspect score distributions, to explore the ability of this population. However, as we will see, by this method alone we were not able to establish cut-offs between putative levels. Which is why we needed the judgemental data.

Judging items and tasks

Nine experts were chosen to participate in the judgements exercise. Four were closely involved in the Year 12 examination reform as members of the English team. One was leader of the Year 10 team for English, one was responsible for the production of the Joint University Entrance Exam for English, one was responsible for the production of the current School-leaving examination for English, one was chief examiner for English at the State Foreign Languages Examination Board and a former member of the Accreditation Committee for Foreign Language Examinations and one was a British Adviser responsible for the development of the in-service course on examination awareness, intended to accompany the English examination reform. All had extensive teaching experience in Hungary, at various levels and because of their responsibilities could be expected to have an authoritative internalised concept of the levels Basic, Intermediate and Advanced. Certainly it would be impossible to find anybody in Hungary with a better idea of what these terms might mean. In addition, all were familiar with the Council of Europe's Common European Framework of reference.

Although we had gathered empirical pilot data on 5 Listening tasks, 12 Reading tasks and 8 Use of English tasks, we decided not to include the Listening tasks in this standard-setting exercise because of the logistical difficulty involved in getting all judges to listen to all the tapes. Furthermore, as this was something of an experiment in standard setting, we did not wish to overburden judges, at least until we were sure that the procedures would be fruitful and until we had an estimate of how long the judgements would take.

We therefore decided to use only the Reading tasks (including the anchor task) and the two anchor Use of English tasks. These latter were chosen for two reasons. One was that we had separately studied one of the Use of English anchor tasks – the sequencing task – in a separate study of sequencing as an item type (see Alderson, Pércsich and Szabó, 2000) and it seemed useful to gather extra data on that task. In addition, that study and an inspection of both anchor tasks suggested that the tasks might be measuring Reading as much as Use of English and we thus judged it useful to see how judges reacted to them in terms of level of difficulty (recall that Use of English tasks are supposed to be at

Advanced level only, not Basic or Intermediate). If the standard setting exercise proved useful, we would likely extend it at some point in the future to include samples of students' writing, and to explore the feasibility of gathering judgements on listening tasks as well.

Materials for the exercise included extracts from the relevant sections of the Common European Framework, an instruction sheet setting out and describing the judgement procedures, copies of the 15 Reading and Use of English test tasks to be evaluated, and copies of the self-assessment statements developed by a European Union project known as DIALANG, which is closely based on the CEF. Appendix 16.1 presents the judgement procedure sheet.

- Procedure 1 asked judges to decide which of the three Hungarian levels – Basic, Intermediate and Advanced – the 15 test tasks were at.
- Procedure 2 required them to re-familiarise themselves with the CEF and to decide at which of the six Council of Europe (CoE) levels each test task was.
- Procedure 3 involved the judges in categorising each DIALANG self-assessment statement into one of the 6 CoE levels (Appendix 16.2).
- Procedure 4 had two parts. In each, the judges had to think of a barely adequate candidate at Intermediate level – somebody who just made the level – and to decide for each item on each test task (124 in all) what the percentage chances would be of that candidate getting the item correct. Then in the second part, they were asked to look at the test task overall, and decide what raw score a Barely Adequate Intermediate candidate would get on that task. These judgements were then to be repeated for a Barely Adequate Advanced candidate.

Two weeks after these procedures had been completed and returned, a fifth procedure (Appendix 16.3) was given to the same judges in preparation for the meeting at which the standard-setting exercise would be discussed.

- Procedure 5 used the modified Angoff procedure developed by Kaftandjieva, Verhelst and Takala in the DIALANG project (Kaftandjieva et al, forthcoming). In this judgement procedure, the judges were asked to decide for each item, on the same Reading/ Use of English test tasks, which items a Barely Adequate Intermediate candidate would get right, and for the same items which ones a Barely Adequate candidate at Advanced level could get wrong, and still be considered an Advanced level student. The aim of presenting this fifth procedure was two-fold: both to experiment with another standard-setting procedure which looked simpler than Procedure 4, and to check on the comparability – reliability, if you like – of the judgements on the two procedures, which ought to be at least roughly similar. Results are presented below.

The experience of the standard setting exercise was then discussed at a meeting attended by eight of the nine judges. At this meeting, judges discussed their responses to Procedure 5 in pairs, and then in plenary, and compared the process of doing Procedure 5 with that of Procedure 4. They were then presented with the results of Procedure 3 (Appendix 16.4) and asked to compare their own responses with the overall results. After a brief discussion of how they had approached this procedure and what they thought of the results, they were then given the results of Procedures 1 – 4 (see Appendix 16.5) and their own responses were returned to them for comparison with the group's results. Again, their impressions of the results were discussed, and views on the value and difficulty of the whole exercise were exchanged and recorded. Inevitably, data relating to Procedure 5 could only be calculated after the meeting.

Results

Raters' comments

In the debriefing session, judges felt unanimously that they had neither needed nor wanted training or a familiarisation session before making the judgments. They preferred to make their own judgements uninfluenced by other people or any familiarisation, and they said they were all sufficiently familiar with the Council of Europe documents and scales, or had had ample opportunity to re-read them to refresh their memories, for there not to be a problem.

One criticism of the Framework was that it tended to focus on text difficulty, rather than task difficulty, whereas when judging items they had, quite rightly, considered task difficulty. On Procedures 1, 4 and 5, at least one judge felt that there were four things to keep in mind when judging task levels:

i) what an ideal Intermediate student ought to be capable of
ii) what is possible for Hungarian learners some time in the future
iii) where Hungarian learners are right now, as shown by the pilot data
iv) what one's own students were capable of.

It was felt that if other judges had also varied in this way, this might account for some differences of opinion among judges, and even internal inconsistency for any one judge. It was also felt that disagreements on specific tasks may have been due to the different extent to which judges might have been influenced by the quality of the test tasks, like less than ideal layout (test task 10), or by text difficulty rather than actual student responses.

All judges said that on Procedure 3 they had not referred back to the Council of Europe scales, but had sufficiently internalised the scales to be able to rate the statements with confidence. By the time they came to do Procedure 5, they felt almost excessively familiar with the Framework and with the tasks. Some judges said that they might have been influenced by their memory of previous judgements (despite the intervening two to three weeks!), by their irritation with the judging procedure, their knowledge of item difficulties, and their impatience. Surprisingly, some judges preferred Procedure 4 – the use of percentage estimates – to Procedure 5, and some felt that Procedure 5 was easier simply because they were by now so familiar with the test tasks and items. Despite our expectations, there was no strong feeling that Procedure 5 was simpler – opinion was divided. Some judges preferred to make global judgements about test tasks – Procedure 4b – rather than making judgements item by item (Procedure 4a), feeling that it was hard to predict which exact items a student would get right or wrong, even if they had a good idea of what score students would get on each test task. Basic test tasks were hard to judge, as they were too easy, and at least one judge felt that it was very easy to decide which items an Advanced student could get wrong (Procedure 5b), since most test tasks were too easy for Advanced students anyway.

One judge consciously tried to be analytic in her judgements, but felt that this was less 'accurate' than relying on an intuitive feel for the test tasks and items. She felt she ought to be more analytical but found it both difficult to be so, and less satisfying. At least one judge reported that when trying to be analytical, she had spent up to 10 minutes considering each item! Interestingly, judges had been asked to keep a record of how long they had taken over the first four judgement procedures, and results varied from six hours to over 25 hours!

All judges reported finding the experience challenging, but very interesting and informative. They were very keen to see how their results compared with others' and with the average, and they would have liked even more detail on how they differed from each other, and on what the spread of results was. They considered the standard-setting procedure a valuable procedure, even if tedious at times. They also felt that such an exercise would be very useful for item writer training.

Variation in judgements

As expected, expert judges differed in their opinions about the difficulty of the various test tasks. Appendix 16.6 gives the details of the differences of opinions for Procedures 1, 2 and 3. The differences in Procedures 4 and 5 were necessarily even greater.

What is one to make of such differences? At one level, we can simply say that these judgements are not to be trusted, because of the variation. At another level, we can say that differences are legitimate, since we are necessarily dealing with subjective matters, and different experts will have different experiences. If differences are legitimate, we simply take the average, on the grounds that more are better than fewer, and that there are no other better ways of resolving differences. In fact, in all the results reported below, we take the median (not mean) judgement on each task to represent the average judgement of our nine expert judges.

However, one possible consequence of these differences, to which we will return, is that if we fail to find convincing evidence on which to establish a strong relationship between expected and actual difficulties of tasks, we will be bound to question the value of such expert judgements, possibly even where they agree with each other. This leaves us with the empirical difficulties alone on which to base decisions about levels.

Comparison of the results of Procedure 5 with Procedure 4

As mentioned above, Procedure 5 was used both in order to try out a potentially simpler procedure, and to enable some estimate of the reliability of the standard-setting process, although the difference in the procedures was acknowledged to make the estimate of test-retest reliability problematic. Moreover, the results for Procedure 5b were problematic in that virtually every task was seen as too easy for Advanced students, and therefore many judges noted that Advanced students would get maximum scores – ie no items wrong.

Appendix 16.7 gives the detailed comparison, test task by test task, of the results of Procedures 4 and 5. Table 16.1 below summarises the results.

Table 16.1: Correlation of Procedures 4 and 5

	Intermediate	**Advanced**
Procedure 4b: 5a	**.93**	**.94**
Procedure 4a: 5a	**.73**	**NS**
Raw score from Procedure 4a: 5a	**.95**	**.95**

If we compare the score for each task arrived at in Procedure 4b, where a score for a Barely Adequate Intermediate candidate was estimated, with the results of Procedure 5a, aggregated from item level judgements to create task-level estimated scores, the result is a Spearman rank order correlation of .93. For Advanced students, the correlation is .94.

We can also compare the score resulting from the item level probability estimates (Procedure 4a) with the score resulting from Procedure 5a (Intermediate). The result is a Spearman rho of .73, but for Advanced students, there is no significant correlation, probably due to the high mean scores, as mentioned above.

We can also relate the results of Procedure 5 to the results of Procedure 4, by calculating the raw score for each task implied by the percentage scores arrived at by estimating percentage chances of success on each item (thus for Task 8, a score of 24% on 15 items is 3.6, rounded up to an estimated score of 4, and for a 10-item test, an estimated score of 73% is an estimated raw score of 7). For Intermediate students the rank order correlation is .95, and for Advanced students, also .95.

It would appear that the results of the standard setting are adequately reliable, in terms of consistency among different judgement procedures.

Item Writer predictions

Table 16.2 below shows how the item writer predictions compare with the standard setting judges:

Table 16.2: Comparison of item writer prediction and judges (Procedure 1)

Judges Item writer	Levels	Basic	Intermediate	Advanced
	Basic	4	2	
	Intermediate		3	1
	Advanced	1	1	2

In 5 out of the 14 cases, there are disagreements, but in the remaining 9 cases, the prediction agrees with the judges. The most radical disagreement was on one of the anchor Use of English tasks, which the item writer had intended to be Advanced (3), but which judges considered to be Basic (1).

Ultimately, however, what matters is not what item writers predict or what judges judge, but the empirical difficulty levels of test tasks. In Table 16.3, we present the results of comparisons of judgements of level by item writers and expert judges with the empirical data from the piloting.

Table 16.3: Comparison of judgements with task difficulties

	Range FV	Mean FV	Range Logits	Mean Logits
Item writer				
Basic	29% – 83%	69%	-2.62 to +0.39	-1.75
Intermediate	20% – 65%	43%	-.83 to +0.71	+0.03
Advanced	24% – 62%	39%	-.64 to +1.89	+0.34
Rho Correlation	.62		.67	
Judges				
Basic	62% – 81%	72%	-2.62 to –0.64	-1.85
Intermediate	26% – 83%	50%	-2.30 to +0.60	-0.39
Advanced	20% – 44%	32%	-0.51 to +1.89	+0.70
Rho Correlation	.75		.73	

If we compare item writer predictions with facility values, we find that for 'Basic', they range from 29% to 83%, for 'Intermediate' from 20% to 65%, and for 'Advanced' from 24% to 62%. The difference between mean difficulty for Intermediate and Advanced is however minimal (43% and 39%). The rank order correlation of item writer predictions with facility values is .62.

For the judges, the relationship of judgements with facility values is somewhat better: for 'Basic', the facility values range from 62% to 81%, for 'Intermediate' from 26% to 83%, and for Advanced from 20% to 44%. Mean facility values decrease steadily, as one would predict, with increased judged difficulty (72% – 50% – 29%). Judges seem better than item writers at predicting facility values, especially for Basic and Advanced tasks. The rank order correlation of judges' estimates of level with facility values is also better, at .75.

If we now compare item writer predictions with calibrated logit values, we find that for 'Basic', they range from -2.62 to +.39, for 'Intermediate' from -.83 to +.71, and for 'Advanced' from -.64 to +1.89. There is little difference between 'Intermediate' and 'Advanced' in their mean logit difficulties. The rank order correlation of item writer predictions with logit values is .67.

For the judges, there is overlap in calibrated difficulty across judged levels: for 'Basic', the logit values range from -2.62 to -.64, for 'Intermediate' from -2.30 to +.60, and for 'Advanced' from -.51 to +1.89. However, the mean empirical logit difficulty increases

steadily with judged difficulty level, and the rank order correlation of judges' estimates of level with logit values is .73.

Judges seem better at predicting empirical difficulty than item writers, but there is still considerable variation in 'accuracy'. It is clear that the empirical difficulty of test tasks does not always correspond to expert judgements about levels. Empirical difficulties are essential before any statement can be made about the 'true' level of any test task.

Comparison of the various standard setting procedures

Procedure 3 was an attempt to see how good the judges were at identifying Council of Europe Framework levels. The data are presented in Appendix 16.4, and the correlations of our judges with the original calibrated Council of Europe levels was an impressive .97. This was even better than the correlation of the DIALANG calibrations with the Council of Europe, where the rank order correlation was .85 (Kaftandjieva, 1999). Interestingly, our judges also correlated at .85 with the Finnish self-assessment data. It would appear that we can have confidence in the ability of our judges to estimate Council of Europe levels, at least of self-assessment statements.

Table 16.4 below shows how the judges' categorisation of the 15 tasks into three Hungarian levels corresponds to their categorisation of tasks into the Council of Europe's six level Framework.

Table 16.4: Judges' Hungarian levels compared with Council of Europe

		PROCEDURE 2				
		A1	A2	B1	B2	C1
PROCEDURE 1	Basic	1	4			
	Intermediate			5	2	
	Advanced				1	2

From this we can see that Basic level is either A1 or A2, Intermediate is mainly B1 but two Intermediate tasks were judged at B2, and Advanced tasks are either C1 or B2. No tasks were rated at C2. Not surprisingly the Hungarian levels cover a wider range than the Council of Europe levels, or, to put it another way, these tasks vary in their level as characterised by the Council of Europe Framework. They are not homogeneous.

How do the judgements of the Council of Europe levels relate to logit values and facility values? Tables 16.5a and b present the data.

Table 16.5a: Relationship between Council of Europe levels and calibrated values of test tasks

Level	Range of logits	Mean logit
A1	-1.64	-1.64
A2	-2.62 to -0.64	-1.90
B1	-2.30 to +0.39	-0.495
B2	-0.83 to +0.60	-0.24
C1	+0.71 to +1.89	+1.30
Spearman rho	+0.73	

Clearly there is overlap in difficulty across test tasks at supposedly different levels, although the relationship is in the expected direction.

For facility values, the data are similar, as follows:

Table 16.5b: Relationship between Council of Europe and facility values

Level	Range of FVs	Mean FV
A1	76%	76%
A2	62% to 81%	71%
B1	29% to 83%	54%

B2	**26% to 65%**	**45%**
C1	**20% to 24%**	**22%**
Spearman rho	**+0.76**	

The overlap of empirical difficulty across levels is considerable, which makes it difficult to arrive at cut-off scores for each level, at least based on these test tasks.

Procedure 4 was designed to facilitate the identification of cut-off scores by getting judges to estimate the performance of barely adequate candidates at both Intermediate and Advanced levels, both item by item, and by total test task score. The rank-order correlation between the two methods, item by item and total task score, is .77 for Intermediate, but it is non-significant for Advanced.

We can calculate possible cut-scores for each task, by taking the judged probability of barely adequate candidates getting an item correct, summing these probabilities, for each item in a task, and calculating, for each judge, the average probability of success by barely adequate candidates on a given task. We can then average these across judges to arrive at a mean 'cut-score' for each task. We can then compare these 'cut-scores' with the tasks' judged Hungarian (Procedure 1) and Council of Europe (Procedure 2) levels. The results are shown in Table 16.6.

Table 16.6: Probable cut-scores for barely adequate candidates, by judged task level

	BA Intermediate		BA Advanced	
	Range	Mean	Range	Mean
Basic	69% – 79%	74%	90% – 98%	94%
Intermediate	33% – 69%	55%	66% – 97%	83%
Advanced	24% – 29%	27%	61% -64%	63%
Rank order rho between Procedure 1 and 4	.92		.81	
Council of Europe				
A1	74%	74%	94%	94%
A2	69 to 79%	74%	90 to 98%	94%
B1	52 to 69%	60%	79 to 97%	87%
B2	29 to 51%	38%	61 to 81%	69%
C1	24 to 29%	27%	63 to 64%	64%
Rank order rho between Procedure 2 and 4	.94		.82	

Impressive though these correlations are, we cannot rule out the possibility that the judges were influenced by their Council of Europe judgements when estimating the probabilities for each item. Nevertheless, it does seem as if we are getting useful information here. We can estimate cut-scores for Intermediate and Advanced candidates on tasks at different levels. The next step would be to apply these cut-scores to the data obtained during the piloting to see the effects on our pilot population. Table 16.7 combines Tables 16.3, 16.5 and 16.6 to compare probable cut-scores with actual task difficulties.

Table 16.7: Comparison of actual facility values with judged mean cut-score for Intermediate and Advanced

	Judged Mean Intermediate Barely Adequate Score	Judged Mean Advanced Barely Adequate Score	Actual Facility Value
Basic	74%	94%	72%
Intermediate	55%	83%	50%
Advanced	27%	63%	29%

A1	74%	94%	76%
A2	74%	94%	71%
B1	60%	87%	54%
B2	38%	69%	45%
C1	27%	64%	22%

What we see here is a degree of correspondence between the estimated cut-score for Intermediate students, and the actual facility values for tasks at the three levels. Of course, the cut-scores for Advanced students are considerably higher than the mean facility values: advanced students have to do very well on test tasks to be considered 'Advanced'. However, even to be considered 'Intermediate', students will have to score at roughly the mean of the population in order to 'pass'. In other words, roughly half the population would not be considered to be 'Intermediate'. This presents considerable problems for decision-makers, since if there are only two levels of exams, Intermediate and Advanced, many students will fail, unlike in the current examination, as we have seen. We will return to this important point later, but before then, we need to explore the data further, to see how robust this apparent finding is.

First, let us consider the issue test task by test task, since Table 16.7 above presents aggregated data. If we compare the score a barely adequate candidate would get on any test task with the mean difficulty of that task – using mean facility values – we can gain some idea of the ability of the population in comparison with the task's difficulty. This enables us to estimate the consequences of setting an exam level at a given 'height', for each test task.

To illustrate this, consider Appendix 16.5, and take, for example, Task 1. The BA (Barely Adequate) Intermediate score is either 33% or 29%, depending on the method used, but the mean difficulty was 20%. This means that this task would be 'failed' by a substantial proportion of the population. Conversely, Task 4, estimated as being Intermediate by both item writers and judges, has a cut-score of 43% or 52% for barely adequate intermediate candidates yet its facility value was 54%. In other words, the proportion of the population 'passing' will vary from task to task, as empirical difficulties do not correspond to judged difficulties. This means that cut-scores, or pass marks cannot be set in advance, and will need to be calculated for the particular set of tasks on any given examination. Once again, this argues for multi-level examinations, not simplistic two-way distinctions. It also argues, we believe, for challenging the notion of 'pass marks' in general. We will return to this issue in our conclusions.

However, note that in the above example, Task 4 was in Booklet 1, whereas Task 1 was in Booklet 3 and we know that the population for Booklet 3 was weaker than that for Booklet 1 (Chapter 8). This means that we need estimates of test task difficulty which are independent of the ability of the sample of the population taking the task, which the above data are not. The difficulty in interpreting these figures emphasises the need for population-free item parameters, and task-free person ability measures. These we have, in our calibrated logit scores for candidates, and calibrated logit item difficulties, and this we shall explore in the next section.

Calculation of ability levels

Since we used anchor items to calibrate item difficulties using IRT (the program BigSteps), we were also able to calculate each person's ability estimate in logits, on the same scale. This enables us to compare the ability of individuals in our population, even if they took partially different test booklets.

We calculated an overall ability estimate for each person, using Reading, Use of English and Listening as anchor items, and we use this estimate in the next chapter (Chapter 17)

as the best overall estimate of language proficiency. We also calculated separate ability estimates for Listening and Reading/ Use of English. However, in this chapter we report an ability estimate for Reading alone, removing Use of English tasks. Recall that two of the tasks judged in the standard-setting exercise were Use of English tasks. Since these were the anchor tasks, and Reading/Use of English ability estimates were calculated based on all Use of English tasks – which we did not include in this standard-setting exercise – we omit the two Use of English tasks from the analysis in this section, and confine the analysis and discussion to the 13 Reading tasks (including the CITO anchor), and a calibrated reading ability based on performance on these tasks.

First, in IRT it is usual to estimate the extent to which the items have measured the ability of the population, by comparing item difficulty estimates with candidate ability estimates, on the same logit scale. Figure 16.1 below gives this data for Reading.

```
MAP OF PERSONS READING ABILITY AND READING ITEMS

MEASURE                                  |                                  MEASURE
         ---------------------  PERSONS-+-  ITEMS  ---------------------
    5.0                        .######## +                                    5.0
                                         |
                                       . |
                                         |
    4.0                                .# +                                    4.0
                                      ## |
                                      .# |
                                     .## |
    3.0                            .##### +                                    3.0
                                     .## | X
                                   .##### | XXX
                                 .######## | XX
    2.0                            .##### + XXXX                               2.0
                                 .######### | XXX
                               .########## | XXXXXXXXXXX
                             .############ | XXX
    1.0                        ########### + XXXX                               1.0
                             .############ | XXXXXXX
                        .######################## | XXXXXX
                               .################ | XXXX
     .0                ######################### + XXXXXXXXXX                    .0
                           .################### | XXX
                              ################## | XXXXXX
                                   ############ | XXXX
   -1.0                              .###### + XXXX                           -1.0
                                   .#### | XX
                                    #### | XXX
                                   .### | XXXXXX
   -2.0                             .## + XX                                   -2.0
                                     .# | XXXX
                                      . | XXXX
                                     .# | XXX
   -3.0                               . + XX                                   -3.0
                                      . | X
                                      # | X
                                        |
   -4.0                                 + X                                   -4.0
                                      . |
                                      . |
                                        |
   -5.0                               . +                                    -5.0
         ---------------------  PERSONS-+-  ITEMS  ---------------------
 EACH '#' IN THE PERSON COLUMN IS   4 PERSONS; EACH '.' IS 1 TO   3 PERSONS
```

Figure 16.1

There is a trend for items to be somewhat more difficult than the people are able. Nevertheless, this map is reassuring that the majority of items were at a suitable level of difficulty, and that virtually all students are measured appropriately.

We know that the reading ability of this population varies enormously – see Figure 16.1 above, and Figure 16.2 and Table 16.8 below.

Table 16.8: Calibrated Reading ability

Number of candidates	mean logit score	Standard deviation	Minimum	Maximum
939	.5923	1.5107	-4.45	4.78

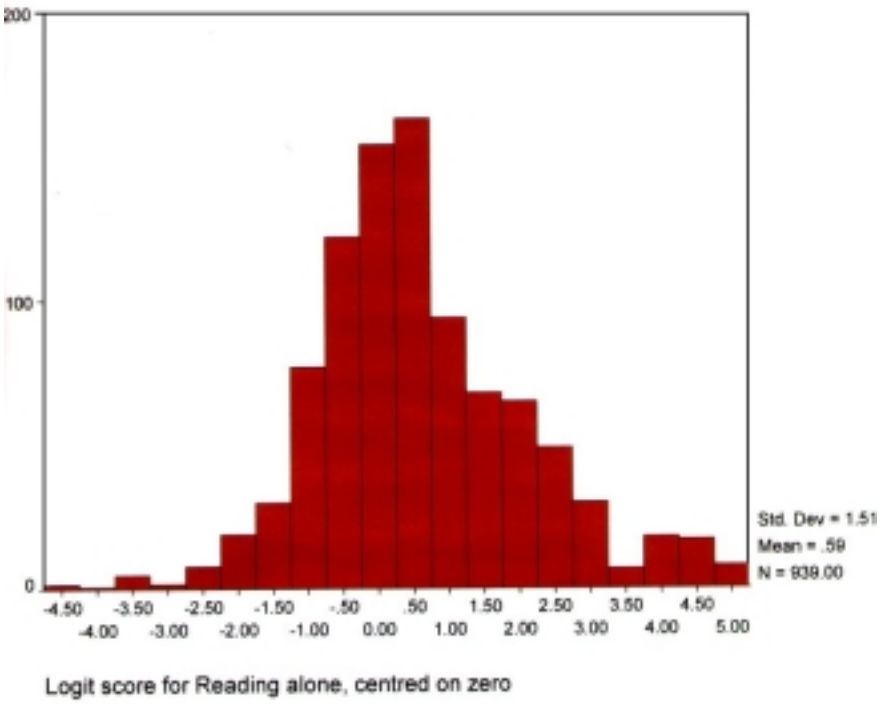

Figure 16.2

Clearly this population is very heterogeneous with respect to reading ability. The question is: what level or levels are they at? Which students are Basic, which are Intermediate, and which are Advanced? Clearly students are spread out in ability, but where should we place the borders between the three 'levels', and how can we characterise students at various points on the scale?

Recall that we have CITO anchor items whose difficulty is 'known'. Both Listening and Reading anchors are thought to be at level A2 in the Council of Europe Framework. We will look at the empirical calibrated values for Listening and Reading separately. The 10 Listening anchor items range in logit values from -2.05 to -0.62, with a mean of -1.02. The 10 Reading anchor items range in logit values from -2.29 to 1.03, with a mean of -0.826. If we plot these values for the anchor items, supposedly at A2, on a distribution of candidate ability scores, we find the following (Table 16.9).

Table 16. 9: Comparison of Reading ability with difficulty of anchor items (A2)

CITO	Range	Mean	Population below minimum	Population above maximum	Population within A2	Population above mean
Listening	-2.05 to -0.62	-1.02	12.5%	39%	49%	45%
Reading	-2.29 to 1.03	-0.826	2%	32%	66%	87%

For Listening, we find that 12.5% of the population score less than -2.05, and 39% score higher than -0.62. In other words, 48.5% of the population fall within CoE level A2. 44.6% of the population has an ability estimate higher than the mean value of -1.02.

For Reading, the anchor items have a wider range of difficulty, and are overall somewhat harder than the Listening items. Thus, we find that a greater proportion of the population (66%) can be considered to be at A2 in ability, but that 87% are above the mean ability for A2.

These results are difficult to interpret. On the one hand, they indicate that only 32% of the population have a Reading ability higher than A2 – which our judges consider to be Basic! On the other hand, only 13% have a Reading ability level below the average required for A2. However, the judges considered the CITO anchor Reading items to be

at B2, and Intermediate in Hungarian terms. That would radically change the view of the reading ability of this population.

For Listening, the dilemma is perhaps not so sharp, but still exists: 49% of the population are at A2, and only 45% are above the mean for A2. If A2 truly is 'Basic' somewhere between 39% and 45% ONLY of the population would be considered at Intermediate or above. Most would fail an Intermediate Listening test.

The problem, clearly, lies in the range of difficulty spanned by items intended to be at A2 (or judged to be at B2). Once more, we are confronted with the problem of reconciling empirical difficulties with intended difficulty.

However, whether these anchor items are A2 or B2, this does not tell us where to put the boundary between Intermediate and Advanced, and it certainly does not tell us how to classify those proportions of the population that fall outwith the limits of the anchor items. In particular, are those who achieve the mean and above for A2 items already Intermediate students or are only those scoring above the maximum anchor item values Intermediate students. At what point do Intermediate students, however defined, become Advanced students? These dilemmas clearly apply for both Listening and Reading results.

To attempt to address these questions, we need to turn back to our judgemental data, and since we did not gather judgements on the Listening tasks, we will confine the discussion that follows to the Reading tasks alone.

Reading ability and Reading task difficulty

We need to be able to decide what the logit ability score would be which an adequate candidate could be expected to get, for any task. In effect, we need to combine empirical data with judgements. Since empirical candidate ability scores and actual item difficulties are calibrated on the same logit scale, we can compare candidates' performance with task difficulties.

First, we select only those items which judges felt (on Procedure 5) that candidates had to get right in order to be considered Intermediate or Advanced, respectively. We then look up the actual logit difficulty of each selected item, and calculate the mean logit score, for each task. That gives a logit score, which a candidate would have to achieve, to have reached a barely adequate level on that task. We can then aggregate results across tasks – to arrive at a more reliable or representative cut-score regardless of task variation, since we know that tasks vary in difficulty. And we can do this regardless of judged or intended task level.

Table 16.10 shows the mean difficulty for all reading items, regardless of judgements, and the mean value of those items that judges thought an Intermediate candidate ought to get right.

Table 16.10: Difficulty of Reading items, in logits

	All Reading Items	'Intermediate Items'
Number of items	110	62
Mean	-.9420	-1.6082
Standard error of mean	.1435	.1678
Standard deviation	1.4705	1.3211
Variance	2.1624	1.7453
Range	6.03	5.13
Minimum	-4.18	-4.18
Maximum	1.85	.95

Table 16.10 shows that the items that Intermediate students must get correct is an easier subset of all reading items: in other words, not surprisingly but reassuringly, there are

empirically difficult items which judges do not consider students have to answer correctly in order to be considered to be at Intermediate level.

If we then take these data, and plot student ability scores against them (Table 16.11) we find that 95% of this population would score above the judged cut-score for Intermediate level items. 35% would be considered to be beyond the maximum cut-score for Intermediate students, and 65% of the population falls within the range of possible cut-scores, depending on which items one considers to be the minimum requirement for an Intermediate student.

Table 16.11: Reading ability levels
Mean judgement Procedure 5a (Intermediate minimum) compared with estimated reading ability

	Range of difficulties of 'Intermediate' items	Mean item difficulty	Population below minimum	Population above maximum	Population within Intermediate	Population above mean
Intermediate Reading	-4.18 to .95	-1.6082	.3%	35%	65%	95%

The mean logit value for Reading tasks that an Intermediate student should get correct is -1.6082. Only 5% of the population score less than this (and therefore 95% score higher). What these data appear to show, if we can believe the relationship between judgements and logit scores, is that at least about 35% of this population can be considered to be at the minimum required level to be considered Intermediate level, and possibly as many as 99.7%!

These data contrast with those of the CITO anchor Reading items, and the interpretation of the reading ability of this population varies enormously, depending on which items one considers to be criterial. Remember that the judges considered the CITO anchor Reading items to be at B2, not A2, although they judged its Hungarian level to be Intermediate (Appendix 16.5).

In short, we have considerable difficulty using these data in arriving at a definition of any level, let alone three different levels.

Conclusion

In this chapter, we have endeavoured to develop procedures that might help us to establish the level(s) of achievement of the pilot population, either in the undefined terms of the Hungarian system – Basic, Intermediate, Advanced – or in terms of the Council of Europe levels as set out in the Common European Framework of reference. We gathered expert judgements from colleagues who could be expected to have internalised, if only intuitively, notions of levels of difficulty. We also explored empirical methods of scaling task difficulty and candidate ability. The latter were based upon Item Response Theory, and used anchor items to calibrate the various tasks in our pilot booklets. Obviously the difficulties we established empirically ARE the difficulties of our tasks. The questions are: how can such difficulty levels best be characterised and how can they be predicted? Is it possible to accommodate the evidently wide range of ability of the pilot population, and the wide range of difficulty of tasks piloted, within a scheme of two or even three levels, as currently required by the Ministry of Education? How might such levels correspond to internationally recognised levels, at which the Ministry is said to be wishing to aim?

We found that expert judges were better at predicting empirical difficulties than were item writers. Their judgements appeared to be reliable, and consistent across different standard-setting procedures, and the order of difficulty they predicted corresponded broadly to the increase in empirical difficulty . However, there was considerable variation in difficulty across items and test tasks, such that it was difficult, if not impossible, to say with confidence that a given task was indeed at a given level. It also proved impossible to identify satisfactory cut-scores for performance, either on individual tasks or on a

whole test battery, which would provide convincing evidence that candidates had attained a given level. Above all, however, it became clear that however levels are defined eventually, the impact of pass and fail decisions will need to be considered very carefully. At present, the examination has no currency, because virtually all students 'pass'. If the examination is to have currency, either some students will have to 'fail' or a different system of reporting results will be required. The setting of levels cannot be determined in isolation from a consideration of pass rates, if the notion of passing and failing is retained, in whatever form.

It is clear that tasks will vary in difficulty, despite the intentions of item writers, or the opinions of expert judges. Thus the true level of an item or task or indeed of a whole test can only be determined post-hoc: by analysing the results.

Therefore, either items and test tasks will need to be pre-calibrated before the live examination is administered (so that their empirical difficulty is known), and items and tasks will have to be taken from item/task banks. Or examinations results can only be issued once the performance of the population has been analysed – a practical and logistical problem.

The alternative is to abandon the notion of two levels, and 'passing and failing', altogether, to abandon the 1-5 scale, and to issue examination results on a scale of difficulty – preferably one that encompasses the heterogeneous nature of the ability of the population, for example from 1 to 100. Users of examination results would then make their own decisions about whether candidates had reached the level appropriate for the purposes for which they were selecting candidates, be that employment, university entrance, or whatever. It is clear that in order to cater for lower achievers who can still perform in English, but at low levels, it will be important to give students who achieve Council of Europe levels A1 or A2 recognition of what they have in fact achieved. The proposed 1 to 100 scale would allow precisely that. It should be noted that the current law for érettségi reform provides for such a scale to be used on the new examination.

It should nevertheless be noted that the levels achieved by our pilot population were achieved without any form of preparation on the part of the students or their teachers for this pilot examination. The task types and indeed the nature of the examination itself was completely unknown to them before they actually took the tests. We can confidently expect that, given adequate preparation, students' levels of achievement will rise, since there is abundant evidence elsewhere that students' performance improves once they become familiar with what is required of them on examinations. This, of course, makes it even more important that levels, standards and 'pass rates' are determined empirically, and are not pre-determined.

Appendix 16.1

Standard Setting for the New School-Leaving Examination in English

Dear Colleague,

Many thanks for agreeing to take part in this study. As you will know, the School-Leaving Examination is to be set at two 'levels': Intermediate and Advanced. In addition, the Year 10 Examination is said to be aimed at a Basic level. The major aim of this study is to provide guidance on what these 'levels' might be, and how they might be defined.

More detailed objectives are:

1) To complement empirical data on student performance with judgemental data on cut-scores.

2) To relate test tasks developed as part of the English Examination Reform Project to the Council of Europe Common Framework.

3) To develop procedures that can be used in future standard- setting exercises.

Methodology

Basically, your task will be to make judgements about the level of difficulty of particular tasks developed and piloted in the Project, in relation to Hungarian levels, and to the Council of Europe's Common Framework.

There are four procedures, as follows: Procedure 1 requires you to classify each piloted reading task as Basic, Intermediate or Advanced. I imagine this will take you no more than one hour. Procedure 2 requires you to classify each reading task at one of the six Council of Europe levels. This may take up to one day, depending on how familiar you are with the CoE levels. Procedure 3 requires you simply to classify a set of reading scales according to the CoE levels. I imagine that this will not require much more than one hour. Procedure 4 is the most complex and unfamiliar, as it asks you to imagine two learners, at two different levels, and to estimate their likely success on each reading task and each item. This could take between one and two days. In all cases, I'd be very grateful if you could keep a record of how long each procedure took you.

Here are the detailed instructions for each procedure.

Procedure 1:

Please read each reading test task carefully, and decide, for each test task, whether it is Basic, Intermediate, or Advanced. Please use your own experience/ judgement or 'gut-feeling' as to what these terms mean, or should mean. Indicate the level you decide in the top-right hand corner of each task.

Procedure 2: Council of Europe's Common European Framework.

The six levels that are defined in this Framework were originally known as:

Breakthrough
Waystage

Threshold
Vantage
Effective Operational Proficiency
Mastery

But these verbal labels have since been replaced (in the 1998 second draft version of the Framework) by lettered/ numbered levels:

A1
A2
B1
B2
C1
C2

Please familiarise yourself with the Framework document – preferably Version Two 1998 (top right hand corner: CC-LANG (95) 5 rev. V) but don't worry if you have an earlier version as they are not that different and the key parts I will copy for you, see below.

Particularly relevant in this respect are Chapters 5 and 8, and the Appendices. However, this is still a lot of information to internalise. Tables 4, 5, 6, 7 and 8 (Chapter 8, pages 128-134) are the best summaries of these levels, and I attach copies of these for your information. The Appendices, page 167 ff, provide examples of scales across the various language skills, and especially relevant to your procedure, which concentrates on the reading tasks, are the scales on pages 175-77 and again I attach copies.

After reading through these documents, and internalising your interpretation of the levels, the procedure is to decide for each reading task whether it is A1, A2, B1 etc. Please write the designated level in the bottom left-hand corner of each task.

Procedure 3: Self-Assessment Scales

A European Commission-funded project, known as DIALANG, has sought to distill much of the information in the Common European Framework, and has developed self-assessment statements from these scales, across the various language skills. I attach copies of the self-assessment scales for Reading.

Your procedure is to decide which of the six levels, A1 – C2, each of these statements represents. You are not required to classify the reading tasks themselves.

DO NOT REFER TO THE ORIGINAL SCALES OF PROCEDURE 2 WHEN MAKING YOUR JUDGEMENTS.

Please indicate your rating in the final column, for each statement from the DIALANG self-assessment scales for reading.

Procedure 4: Cut-scores and Barely Adequate performances

Think of a Hungarian student, in the Fourth Year of Upper Secondary School, who is likely **only barely** to pass the Intermediate level test. Call this student: Barely Adequate Intermediate (BAI).

Think of another student, who is likely **only barely** to pass the Advanced level test. Call this student: Barely Adequate Advanced (BAA).

Now, for each reading test task,

i) indicate against each item what the percentage chances are of, first, a Barely Adequate Intermediate (BAI) student getting the item correct (5%? 40%? 85%. Use any figure you think appropriate), and then, secondly, what the percentage chances are of a Barely Adequate Advanced (BAA) student getting the item correct.

ii) Once you have done this for each item in the first task, estimate what total score a Barely Adequate Intermediate student would get on this task, and then what total score a Barely Adequate Advanced student would get on the same task.

iii) Only then go on to the next reading task, and repeat subtasks i) and ii) above. This should be done for all the tasks in the reading battery.

YOU MAY BE TEMPTED TO CROSS CHECK YOUR JUDGEMENTS AND CONTINUALLY TO REFER BACK TO PREVIOUS ITEMS. THIS SHOULD NOT BE NECESSARY, AS IT IS IMPOSSIBLE TO GET THESE ESTIMATES 100% CORRECT OR TO ACHIEVE 100% RELIABILITY.

PLEASE DO NOT CONSULT WITH ANY COLLEAGUES AT ANY STAGE DURING THIS PROCESS. IF YOU HAVE ANY QUESTIONS ABOUT PROCEDURES, PLEASE CONTACT CHARLES ALDERSON ONLY. WE WILL HOLD A MEETING AFTERWARDS TO DISCUSS THE RESULTS.

The completed procedure sheets should be returned to Charles Alderson **by January 31st at the very latest**

Many thanks for your collaboration. This will prove invaluable in deciding what level our various tasks might be at, and should be at.

Appendix 16.2

Standard-setting for English Examinations Reform Project

Many thanks for agreeing to take part in this standard setting exercise. Please see separate procedure sheet for instructions.

Your name.....
Date.....

Please indicate which of the six levels (A1 to C2) is best characterised by each statement

No	Self-assessment statements	CoE Level
1	I can find specific information in simple everyday material such as advertisements, brochures, menus and timetables.	
2	I can recognise the general line of argument in a text but not necessarily in detail.	
3	I can quickly identify the content and relevance of news items, articles and reports on a wide range of professional topics, deciding whether closer study is worthwhile.	
4	I can recognise familiar names, words and very simple phrases on simple notices in the most common everyday situations.	
5	I can search one long or several short texts to locate specific information I need to help me complete a task.	
6	I can understand and interpret practically all forms of written language including abstract, structurally complex, or highly colloquial literary and non- literary writings.	
7	I can understand short, simple texts written in common everyday language.	
8	I can understand specialised articles outside my field, provided I can use a dictionary to confirm terminology.	
9	I can understand the general idea of simple informational texts and short simple descriptions, especially if they contain pictures which help to explain the text.	
10	I can identify specific information in simple written material such as letters, brochures and short newspaper articles describing events.	
11	I have a broad reading vocabulary, but I sometimes experience difficulty with less common words and phrases.	
12	I can understand standard routine letters and faxes on familiar topics.	
13	I can understand simple instructions on equipment encountered in everyday life – such as a public telephone.	
14	I can understand everyday signs and notices in public places, such as streets, restaurants, railway stations and in workplaces.	
15	I can understand straightforward texts on subjects related to my fields of interest.	
16	I can find and understand general information I need in everyday material, such as letters, brochures and short official documents.	
17	I can understand short, simple messages e.g. on postcards.	
18	I can understand articles and reports concerned with contemporary problems in which the writers adopt particular stances or viewpoints.	
19	I can identify the main conclusions in clearly written argumentative texts.	
20	I can understand very short, simple texts, putting together familiar names, words and basic phrases, by for example rereading parts of the text.	
21	I can understand the description of events, feelings and wishes in	

	personal letters well enough to correspond with a friend or acquaintance.	
22	I can understand short, simple texts containing the most common words, including some shared international words.	
23	I can read correspondence relating to my fields of interest and easily understand the essential meaning.	
24	I can understand short simple texts related to my job.	
25	I can read many kinds of texts quite easily at different speeds and in different ways according to my purpose in reading and the type of text.	
26	I can understand short simple personal letters.	
27	I can follow short, simple written instructions, especially if they contain pictures.	
28	I can recognise significant points in straightforward newspaper articles on familiar subjects.	
29	I can understand any correspondence with an occasional use of dictionary.	
30	I can understand in detail long, complex instructions on a new machine or procedure even outside my own area of speciality if I can reread difficult sections.	
31	I can understand clearly written straightforward instructions for a piece of equipment.	

Appendix 16.3

Standard Setting Procedure 5

This is the final procedure in the Standard setting exercise, I promise!

Please complete this procedure before coming to the meeting on February 29th. If for some reason you are unable to attend the meeting, please mail the completed procedure sheets to:
Charles Alderson,
The British Council
1068 Budapest
Benczur utca 26

There are two subprocedures. The first is to look at each reading task (attached) and decide for each item whether an Intermediate student – a student who is just about an adequate student at Intermediate level – should be able to answer it correctly.
Circle Yes or No.
'Yes' means an Intermediate student should be able to get it right.
'No' means an Intermediate student can get it wrong and still be Intermediate (ie, it is too difficult).

The second sub procedure is somewhat different, and is this time aimed at Advanced students – students who just about make the Advanced level. Again look at each task, and decide for each item which ones a student at Advanced level could answer WRONGLY and yet still be at Advanced level .
Circle Yes or No
'Yes' means an Advanced student can get it wrong, and still be Advanced
'No' means an Advanced student must get it right to be considered Advanced.

Standard Setting Procedure 5a NAME.....

Intermediate Reading
Do you agree that a person with a language proficiency at Intermediate in Reading should be able to answer the following items correctly?

Task 1 A smashing case

Item 1	**Yes**	**No**
Item 2	**Yes**	**No**
Item 3	**Yes**	**No**
Item 4	**Yes**	**No**
Item 5	**Yes**	**No**
Item 6	**Yes**	**No**

Task 2 Girls-only success 'based on selection'

Item 1	**Yes**	**No**
Item 2	**Yes**	**No**
Item 3	**Yes**	**No**
Item 4	**Yes**	**No**
Item 5	**Yes**	**No**
Item 6	**Yes**	**No**
Item 7	**Yes**	**No**
Item 8	**Yes**	**No**
Item 9	**Yes**	**No**
Item 10	**Yes**	**No**

(etc)

Standard Setting Procedure 5b

Advanced Reading
Please mark the items which a person could answer <u>WRONGLY</u>
and you would still be willing to consider him/ her at an Advanced level in Reading

Task 1 A smashing case

Item 1	**Yes**	**No**
Item 2	**Yes**	**No**
Item 3	**Yes**	**No**
Item 4	**Yes**	**No**
Item 5	**Yes**	**No**
Item 6	**Yes**	**No**

Task 2 Girls-only success 'based on selection'

Item 1	**Yes**	**No**
Item 2	**Yes**	**No**
Item 3	**Yes**	**No**
Item 4	**Yes**	**No**
Item 5	**Yes**	**No**
Item 6	**Yes**	**No**
Item 7	**Yes**	**No**
Item 8	**Yes**	**No**
Item 9	**Yes**	**No**
Item 10	**Yes**	**No**

etc

Appendix 16.4

Standard-setting for English Examinations Reform Project

Procedure 3 results

No	Self-assessment statements	Council of Europe Level	DIALANG Level	Median Level on Proc 3
1	I can find specific information in simple everyday material such as advertisements, brochures, menus and timetables.	A2	A1	A2
2	I can recognise the general line of argument in a text but not necessarily in detail.	B1	A2	B1
3	I can quickly identify the content and relevance of news items, articles and reports on a wide range of professional topics, deciding whether closer study is worthwhile.	B2	B2	C1
4	I can recognise familiar names, words and very simple phrases on simple notices in the most common everyday situations.	A1	A1	A1
5	I can search one long or several short texts to locate specific information I need to help me complete a task.	B1	B1	B1
6	I can understand and interpret practically all forms of written language including abstract, structurally complex, or highly colloquial literary and non- literary writings.	C2	C1	C2
7	I can understand short, simple texts written in common everyday language.	A2	A2	A2
8	I can understand specialised articles outside my field, provided I can use a dictionary to confirm terminology.	B2	B2	C1
9	I can understand the general idea of simple informational texts and short simple descriptions, especially if they contain pictures which help to explain the text.	A1	A2	A1
10	I can identify specific information in simple written material such as letters, brochures and short newspaper articles describing events.	A2	A2	A2
11	I have a broad reading vocabulary, but I sometimes experience difficulty with less common words and phrases.	B2	B2	B2
12	I can understand standard routine letters and faxes on familiar topics.	A2	A2	B1
13	I can understand simple instructions on equipment encountered in everyday life – such as a public telephone.	A2	A2	A2
14	I can understand everyday signs and notices in public places, such as streets, restaurants, railway stations and in workplaces.	A2	A1	A2
15	I can understand straightforward texts on subjects related to my fields of interest.	B1	A2	B1
16	I can find and understand general information I need in everyday material, such as letters, brochures and short official documents.	B1	A2	B1
17	I can understand short, simple messages e.g. on	A1	A1	A2

	postcards.			
18	I can understand articles and reports concerned with contemporary problems in which the writers adopt particular stances or viewpoints.	B2	B1	B2
19	I can identify the main conclusions in clearly written argumentative texts.	B1	B1	B1
20	I can understand very short, simple texts, putting together familiar names, words and basic phrases, by for example rereading parts of the text.	A1	A1	A1
21	I can understand the description of events, feelings and wishes in personal letters well enough to correspond with a friend or acquaintance.	B1	B1	B1
22	I can understand short, simple texts containing the most common words, including some shared international words.	A2	A2	A2
23	I can read correspondence relating to my fields of interest and easily understand the essential meaning.	B2	A2	B2
24	I can understand short simple texts related to my job.	A2	A1	A2
25	I can read many kinds of texts quite easily at different speeds and in different ways according to my purpose in reading and the type of text.	B2	B1	C1
26	I can understand short simple personal letters.	A2	A1	A2
27	I can follow short, simple written instructions, especially if they contain pictures.	A1	A1	A1
28	I can recognise significant points in straightforward newspaper articles on familiar subjects.	B1	A2	B1
29	I can understand any correspondence with an occasional use of dictionary.	C1	B1	C1
30	I can understand in detail long, complex instructions on a new machine or procedure even outside my own area of speciality if I can reread difficult sections.	C1	C1	C1
31	I can understand clearly written straightforward instructions for a piece of equipment.	B1	A2	B1

Appendix 16.5

Results of Standard Setting Exercise, Procedures 1 – 4

Task	Item Writer	Proc 1	Proc 2	Logit	FV%	Proc 4 BAI%	Proc 4 BAA%	Proc 4 BAI Score	Proc 4 BAA Score
Task1 (3:5) k=6	2	3	C1	+0.71	20%	29	63	2 33%	4 67%
Task 2 (4:5)k=10	3	2	B2	+0.601	26%	33	66	3 30%	8 80%
Task 3 (3:4)k=10	1	1	A2	-1.91	68%	73	90	6 60%	9 90%
Task 4 (1:3)k=7	2	2	B1	-0.187	54%	52	79	3 43%	6 86%
Task 5 (1:2)k=10	1	2	B1	-2.298	83%	69	88	6 60%	10 100%
Task 6 Anchor k=10	CITO A2	2	B2	-0.826	65%	51	81	4 40%	9 90%
Task 7 (4:3)k=8	3	3	B2	-0.505	44%	29	61	2 25%	4 50%
Task 8 Anchor1 k=15	3	3	C1	+1.891	24%	24	64	2 13%	10 67%
Task 9 Anchor2 k=4	3	1	A2	-0.635	62%	69	98	3 75%	4 100%
Task 10 (4:2)k=5	1	2	B1	+0.394	29%	60	91	3 60%	5 100%
Task 11 (4:4)k=7	2	2	B1	-0.733	49%	62	97	3 43%	7 100%
Task 12 (3:3)k=10	1	1	A2	-2.428	74%	79	94	7 70%	9 90%
Task 13 (3:2)k=8	1	1	A2	-2.624	81%	76	93	5 63%	8 100%
Task 14 (2:2)k=5	1	1	A1	-1.64	76%	74	94	4 80%	5 100%
Task 15 (2:3)k=9	2	2	B1	+0.349	47%	57	81	6 67%	8 89%

Appendix 16.6: Variations in judgements of levels

Procedure 1: Hungarian levels Basic, Intermediate and Advanced

	Median	Spread (number = numbers of judges identifying this level)
Test task 1	Advanced	Intermediate 2, Advanced 7
Test task 2	Intermediate	Intermediate 5, Advanced 4
Test task 3	Basic	Basic 7, Intermediate 2
Test task 4	Intermediate	Intermediate 7, Advanced 2
Test task 5	Intermediate	Basic 4, Intermediate 5
Test task 6	Intermediate	Intermediate 7, Advanced 2
Test task 7	Advanced	Intermediate 3, Advanced 6
Test task 8	Advanced	Advanced 9
Test task 9	Basic	Basic 5, Intermediate 4
Test task 10	Intermediate	Intermediate 9
Test task 11	Intermediate	Basic 3, Intermediate 6
Test task 12	Basic	Basic 7, Intermediate 2
Test task 13	Basic	Basic 7, Intermediate 2
Test task 14	Basic	Basic 7, Intermediate 2
Test task 15	Intermediate	Basic 1, Intermediate 8

Procedure 2: Council of Europe levels

	Median	Spread (number = numbers of judges identifying this level)
Test task 1	**C1**	**B2 4, C1 5**
Test task 2	**B2**	**B1 2, B2 4, C1 3**
Test task 3	**A2**	**A2 8, B1 1**
Test task 4	**B1**	**B1 5, B2 4**
Test task 5	**B1**	**A2 3, B1 6**
Test task 6	**B2**	**B1 4, B2 3, C1 2**
Test task 7	**B2**	**B1 1, B2 4, C1 3, C2 1**
Test task 8	**C1**	**B2 4, C1 5**
Test task 9	**A2**	**A1 1, A2 4, B1 4**
Test task 10	**B1**	**B1 8, B2 1**
Test task 11	**B1**	**A2 3, B1 6**
Test task 12	**A2**	**A1 2, A2 5, B1 2**
Test task 13	**A2**	**A1 3, A2 5, B2 1**
Test task 14	**A1**	**A1 5, A2 3, B1 1**
Test task 15	**B1**	**A2 1, B1 6, B2 2**

Procedure 3 results

No	Self-assessment statements	Median Level on Procedure 3	Spread of Levels+
1	I can find specific information in simple everyday material such as advertisements, brochures, menus and timetables.	A2	A1 2,A2 7
2	I can recognise the general line of argument in a text but not necessarily in detail.	B1	A2 2,B1 6, B2 1
3	I can quickly identify the content and relevance of news items, articles and reports on a wide range of professional topics, deciding whether closer study is worthwhile.	C1	B2 3, C1 5, C2 1
4	I can recognise familiar names, words and very simple phrases on simple notices in the most common everyday situations.	A1	A1 9
5	I can search one long or several short texts to locate specific information I need to help me complete a task.	B1	A2 1, B1 7, B2 1
6	I can understand and interpret practically all forms of written language including abstract, structurally complex, or highly colloquial literary and non- literary writings.	C2	C2 9
7	I can understand short, simple texts written in common everyday language.	A2	A2 7, B1 2
8	I can understand specialised articles outside my field, provided I can use a dictionary to confirm terminology.	C1	B2 3, C1 5, C2 1
9	I can understand the general idea of simple informational texts and short simple descriptions, especially if they contain pictures which help to explain the text.	A1	A1 5, A2 4
10	I can identify specific information in simple written material such as letters, brochures and short newspaper articles describing events.	A2	A2 7, B1 2
11	I have a broad reading vocabulary, but I sometimes experience difficulty with less common words and phrases.	B2	B1 1, B2 6, C1 2
12	I can understand standard routine letters and faxes on familiar topics.	B1	A2 2, B1 7
13	I can understand simple instructions on equipment encountered in everyday life – such as a public telephone.	A2	A1 1, A2 7, B1 1
14	I can understand everyday signs and notices in public places, such as streets, restaurants, railway stations and in workplaces.	A2	A1 4, A2 5
15	I can understand straightforward texts on subjects related to my fields of interest.	B1	B1 7, B2 2
16	I can find and understand general information I need in everyday material, such as letters, brochures and short official documents.	B1	A2 1, B1 8
17	I can understand short, simple messages e.g. on postcards.	A2	A1 3, A2 5
18	I can understand articles and reports concerned with contemporary problems in which the writers adopt particular stances or viewpoints.	B2	B2 8, C1 1
19	I can identify the main conclusions in clearly written argumentative texts.	B1	B1 5, B2 4
20	I can understand very short, simple texts, putting together familiar names, words and basic phrases, by for example rereading parts of the text.	A1	A1 9
21	I can understand the description of events, feelings and wishes in personal letters well enough to correspond with a friend or acquaintance.	B1	A2 2, B1 7

22	I can understand short, simple texts containing the most common words, including some shared international words.	A2	A1 3, A2 6
23	I can read correspondence relating to my fields of interest and easily understand the essential meaning.	B2	A2 1, B1 2, B2 6
24	I can understand short simple texts related to my job.	A2	A2 5, B1 4
25	I can read many kinds of texts quite easily at different speeds and in different ways according to my purpose in reading and the type of text.	C1	B2 2, C1 5, C2 2
26	I can understand short simple personal letters.	A2	A1 1, A2 8
27	I can follow short, simple written instructions, especially if they contain pictures.	A1	A1 6, A2 3
28	I can recognise significant points in straightforward newspaper articles on familiar subjects.	B1	B1 9
29	I can understand any correspondence with an occasional use of dictionary.	C1	B2 1, C1 8
30	I can understand in detail long, complex instructions on a new machine or procedure even outside my own area of speciality if I can reread difficult sections.	C1	B2 1, C1 6, C2 2
31	I can understand clearly written straightforward instructions for a piece of equipment.	B1	A2 2, B1 6, B2 1

+number = numbers of judges identifying this level

Appendix 16.7

Comparison of Procedures 4 and 5

k=number of items, all figures are medians
Figures in brackets (3:5) indicate booklet and part (Booklet 3 Part 5) of the original test

Task	Procedure 4a BAI%	Procedure 4a BAA%	Procedure 4b BAI Score	Procedure 4b BAA Score	Procedure 5a (Inter)	Procedure 5b (Adv.)
Task1 (3:5) k=6	29	63	2 33%	4 67%	2	4
Task 2 (4:5)k=10	33	66	3 30%	8 80%	4	10
Task 3 (3:4)k=10	73	90	6 60%	9 90%	8	10
Task 4 (1:3)k=7	52	79	3 43%	6 86%	4	7
Task 5 (1:2)k=10	69	88	6 60%	10 100%	9	10
Task 6 Anchor k=10	51	81	4 40%	9 90%	6	8
Task 7 (4:3)k=8	29	61	2 25%	4 50%	2	5
Task 8 Anchor1 k=15	24	64	2 13%	10 67%	4	12
Task 9 Anchor2 k=4	69	98	3 75%	4 100%	4	4
Task 10 (4:2)k=5	60	91	3 60%	5 100%	4	5
Task 11 (4:4)k=7	62	97	3 43%	7 100%	5	7
Task 12 (3:3)k=10	79	94	7 70%	9 90%	8	10
Task 13 (3:2)k=8	76	93	5 63%	8 100%	8	8
Task 14 (2:2)k=5	74	94	4 80%	5 100%	5	5
Task 15 (2:3)k=9	57	81	6 67%	8 89%	6	8