Chapter 6

Marking the Pilot Tests

Szabó Kinga, Sulyok Andrea and Charles Alderson

Once the piloting of the various tasks was complete, it was necessary to mark the performances before the data could be processed. In the case of the April booklets for Listening, Reading and Use of English, this was relatively straightforward, although central marking was novel in the Hungarian context. However, in the case of the marking of performances on the Speaking and Writing tasks, it was much more complex, as marking scales and schemes had first to be developed and trialled. In what follows, we deal separately with the marking of the April booklets, and the subjective marking of the productive skills.

Marking of the April 1999 pilots, May 11 and 12

Arrangements were made for marking to take place centrally, in Budapest, to ensure control over the process. Mark schemes and keys were developed in advance, using the suggested keys submitted by item writers along with their items as the basis, and supplementing these after inspection of the items themselves, adding suggestions made by the Editing Committee and others as to acceptable responses. The objective items (multiple choice) were marked by machine, using the facility provided by the software Iteman, candidate responses having been input directly by data processors (see Chapter 7).

The marking took place over 2 days, with 13 markers on the first day, and 7 on the second (it had been anticipated that markers from outside Budapest would not need to attend for a second day, thus saving on time and money). In the event, most markers stayed on until 18.30 on the first day to ensure that the task could be completed in two days. Partly as a result, the second day of marking finished earlier than expected, at 14.00. Thus, the first day lasted too long, and it was concluded that marking sessions must not exceed six hours total on task, plus breaks. A typical day should last from 9.30 to 5 maximum, with at least an hour for lunch.

Marking took the following times:

Listening (in each case only two tasks): 20 scripts per hour per marker
Reading 1 (including the hard task): 10 scripts per hour per marker
Reading 2: 20 scripts in 90 minutes, ie 12/13 scripts per hour per marker
Reading 3: 20 scripts in 90 minutes, ie 12/13 scripts per hour per marker
Reading 4: 20 scripts in 90 minutes, ie 12/13 scripts per hour per marker

The total time taken and the number of scripts marked was as follows:

Table 6.1: Time taken to mark scripts

Paper	Scripts	Total hours	Markers	Time taken
Listening1	241	12.05	12	1 hour
Listening2	267	13.35	13	1 hour
Reading 1	257	25.7	13	2 hours
Reading 2	265	22.08	13	1 hr 40
Reading 3	240ish	20.00	11/6	2 hours
Reading 4	234	18.00	7	2 hrs 58

On the first day, marking took place in two rooms, because of the size of the rooms, the number of markers, and the presence of smokers amongst markers. It would have been better to have had one room big enough for all, and to allow smokers to leave the room on occasion, as necessary.

All markers marked the same papers. Each session started by the 'Chief Examiner' explaining the mark scheme for a given booklet, and asking each marker to mark one script, and ask questions. Markers then exchanged scripts and re-marked. Generally this worked well. Ideally, there would have been a system of sample check marking, as occasionally errors were found in marking or coding. In future, some systematic method needs to be found for sample double-checking of marking.

The Chief Examiner was constantly and continuously available to answer queries, to clarify mark schemes, to decide on what might be acceptable or to call for a group discussion. It is essential that this role be filled by somebody, and the consensus of markers was that this should be a native speaker. Experience shows that it is best if that person does not have to do any marking him/herself, but is available to monitor and to answer any questions. It is essential that the marking team MUST ask the person in charge if they have any questions, especially regarding possibly correct answers, so that all the marking team hears the answer.

The mark scheme was immediately updated when unexpected answers were found. The mark scheme must be updated, and be available for the record in the files of the pilot testing. This also applies to the general instructions – eg: 'Answers in Hungarian are acceptable' (or not).

Mark schemes should be prepared in a user-friendly way - so that they match the answer sheets, and do not merely list correct answers. For ease of marking, the 'circle-the-error' task type, in particular, needs a special transparent overlay (OHP slide) that can be placed over the text, with sufficient space left on the right hand side or beneath the printed task, for item correctness to be indicated. There were serious problems marking the Reading test in Booklet 1, Part Six, as a result of this not having been done.

Since many answer sheets were in the order in which they had been collected during the piloting, it was possible to detect copying in some cases. It is important both to ensure during test administration that copying is impossible but also to develop rules for what to do if the presence of cheating is suspected during marking. Candidates should be told what will happen (papers of both copier and copyee should be cancelled, since it is impossible to distinguish).

The marking of sequencing items presented particular problems, which resulted in the exploration of innovative computer scoring methods - as detailed in Alderson, Pércsich and Szabó (2000).

Markers' comments during marking were especially useful, and some means should be found to record comments on tasks/ items made by markers, for future reference after the statistical analysis is complete.

For example, the Use of English Anchor Task should not have the indefinite article 'a' as an example, since many students seem to have interpreted this to mean 'the letter 'a' must be given if the sentence is incorrect'.

Similarly, it is necessary to reconsider whether, in the same Use of English task, students are to be allowed to tick if a line is correct, since one way of getting answers correct by chance is by ticking many sentences.

Comments volunteered by markers after the marking sessions were complete were as follows:

'I enjoyed Tuesday thoroughly, and if I had known in advance that I would be needed on Wednesday as well, I would have stayed. I learned a great deal about marking, about items and about marker training.'

'Actually, one of my main experiences in all this was how important it is to have somebody to coordinate the marking and to make the final decisions in problematic cases. Tuesday strengthened my belief that central marking is necessary, and that team marking is much better than individual. Before Tuesday I had only known this through common sense and my reading the literature on testing, but Tuesday's marking session made the whole thing somehow real. So I am really grateful for the opportunity.'

'I found the way we did the marking good, I think it was very important that before starting a paper you asked us to mark one script. It was also good to mark the same paper at the same time.'

'I think in the future it will be essential to find an efficient way of double checking - on Tuesday and Wednesday we simply did not have enough time and people for this.'

'I absolutely agree with your comment about how to make mark schemes more user-friendly: the wrong layout made our lives much harder, and caused about 30% of our tiredness towards the end of the day.'

'It was necessary to have a native speaker, thanks for your help, indeed.'

'It would have been useful to read the tasks in advance to see them as a whole (time constraints, I know).'

'I think some re-checking (at random) during correction would have reduced the mistakes made by the markers.'

'The keys should exactly follow the tasks - that's very important - (e.g. in Reading 1, part 4 - the 'correct' lines could be in bold).'

'The markers might be asked later to make notes during correction (e.g. which tasks were found too difficult/easy, were not filled in, the commonest mistakes).'

'The test administrators should feel more responsibility and they should be made to pay more attention to monitoring the students during writing (exact times, correct boxes, writing in 'tick' or 'a') - stricter instructions/rules needed.'

Marking of writing

In the existing School-leaving Examination for English, there are not many guidelines which teachers, who are responsible for marking school-leaving exams, can use when marking

students' written performance. To mark Part A: 'The teachers do the marking themselves, following centrally provided instructions, guidelines and norms' (Fekete et al, 1999:25) and for Part B (Translation from English into Hungarian): 'The assessment of translation is carried out on the basis of generic guidelines (one page) issued by the ministry. They are based on the Hungarian marking system: 5 being the top mark and 1 a failure; however, they define the content only of grades 5 and 1. Those in-between are left to the subjective interpretation of the teachers marking the papers. A top grade is to be given to the translation if the whole text is translated appropriately, the content of the translated text is true to the original and is rendered in good Hungarian. Translation not attempted or totally misunderstood is a fail. When giving the marks in between, the teacher should consider misuse of dictionary, distortions in meaning, poor Hungarian wording and stylistic mistakes.' (op cit: 27)

The marking guidelines provided centrally are very open to individual interpretation. Although teachers working in a school may co-operate when trying to make sense of such guidelines as there are, this does not necessarily happen, and there is in any case no central monitoring of marking standards. In the English Examination Reform Project it was not only decided to standardise the criteria and to develop scales for the assessment of writing but the draft scales were also tried out and we examined how they worked - or did not work. We quickly realised that developing scales was 'easier said than done'.

Setting criteria and describing bands

We worked on the detailed scales during the Manchester training in January, 1999. The team consisted of people trained in testing in general, but they had not done such a task before. The general criteria for assessment had already been laid down in the Specifications, and what was needed now was to give meaning to what performance was expected in different bands. The team decided to think in terms of 10 bands (0-9) per criterion, so that there would not be too many similar bands or too few dramatically different level descriptions. At the time the team viewed the three levels of the future secondary examinations as three stages on a continuum of language proficiency. Nevertheless, level setting had to take place – 'Basic', 'Intermediate' and 'Advanced' had to be defined in real terms.

The *Specifications* also contained ideas on how the levels relate to levels in other types of language examinations. As the Specifications document had been compiled paying special attention to what the Common European Framework expects on different levels we hoped there would be no major catastrophe ahead.

After comparison and discussion of sample marking of scripts, a draft set of criteria was developed, and descriptors were written for these criteria. The four criteria were:

- Task Achievement
- Grammar and Spelling
- Vocabulary
- Organisation.

The wording of the assessment scales was worked on in small groups who from time to time came together to check on progress and to contribute to each others' work as necessary. In the evenings selected scripts were marked using the scales in their current state, which often gave group members further points to look into and work on. At the same time those responsible for different criteria were asked to identify scripts they thought were typical samples of performance in each band.

At the next stage, when draft scales were in a form which was considered a good starting point, everybody was asked to mark copies of the same test script, to see to what extent the scales constrained personal variation in interpretation. Wording was frequently adjusted in

the light of such differences, yet team members often felt they were not always speaking the same language.

The next step was for real: after the Manchester training, team members volunteered to mark and double mark piles of 48 scripts at home. The scripts were to be marked according to the scales (see Appendix 6.1) and markers were also asked to make comments on what shortcomings the scales had, and what problems they had experienced while doing the marking. Markers had to fill in the 'Standardised Assessment Sheet for Writing' below.

Candidate's ID number:	**Assessor's ID number:**	
Criterion	**TASK 1**	**TASK 2**
TASK ACHIEVEMENT		
GRAMMAR AND SPELLING		
VOCABULARY		
ORGANISATION		

After marking 24 scripts, markers exchanged scripts with a colleague. Markers, naturally, did not have access to other markers' assessment sheets. In Chapter 10 we give the details of the comments made by markers, both on the criteria, and on the tasks, and in the same chapter we present a detailed analysis of the data resulting from this marking.

Many problems and questions came up during these rounds of marking. It was quite clear that the marking scales in their present form were a very rough working version that needed much further work. Nevertheless, developing the scales and marking according to them were not in vain as they provided a good starting-point and the feedback from the markers gave extremely important guidance on what and how to improve and change.

Assessing oral performance

The current school-leaving oral examination consists of two tasks. 'Marking is done by the candidates' own teacher, and the oral score forms 50% of the final result.'(Fekete et al, 1999:27). 'On the whole, the two tasks don't seem to provide enough information for the examiner to form an opinion of the candidates' oral competence. In addition, it is obvious that too much depends on the attitude and active participation of the examiner, who is the candidate's own teacher. Most important of all, however, there are no centrally developed exam tasks and grading criteria for the comparative assessment of the candidates' performance, so any judgement of the examiner is only worth as much as any holistic subjective estimate, and may not be valid outside the examination room.' (op.cit., p.34.)

Given the current state of affairs, it was necessary to devise a more satisfactory system of grading students' oral performances. This involved the development and trialling of rating scales, in a similar fashion to the work done developing scales for rating written performance.

Setting criteria and describing bands

As described in Chapter 5, the Speaking tasks were piloted in schools in December 1998, and all performances were videotaped. Altogether 79 students were tested, 54 of them in the paired mode. In January 1999, during the Manchester training, the criteria and scales for assessment were worked on, since by that time we had a relatively large number of sample performances.

We agreed that we would need analytic rating scales to assess these performances. First, we drew up scales for writing as detailed in the previous section, and then tried to use the same principles, as far as possible, for speaking. For speaking we had agreed on the following four rating criteria in the Specifications document:

- overall communication,
- grammar (range and accuracy),
- vocabulary (range and accuracy),
- speech quality.

In Manchester we agreed on the number of levels, on a 0 to 9 level scale, on which we hoped to be able to place performances of all three levels. We were very careful to associate our levels with the ones of the Common European Framework. The next stage was to create band-descriptors using the Common European Framework as well as scales already established by the team working on the reform of the German school-leaving examination. In the end we came up with what was the first draft of the speaking scales. After refining them in a similar process to that described for the development of the writing scales, the second draft was drawn up. (Appendix 6.2) This was then used for the assessment of videotaped performances, after the Manchester training. In addition, sample videotaped performances were identified at each level on the four analytic scales, for reference and future training purposes.

Each candidate performance was assessed independently by two assessors, and they completed a 'Standardised Assessment Sheet for Speaking' (Appendix 6.3) for each performance. Both candidate and assessor data were coded for later analyses of the results. In addition, assessors were asked to give feedback on how they thought the scales and the tasks worked.

In Chapter 11, we give the details of the comments made by markers, both on the criteria, and on the tasks, and we present the details of the empirical analyses, such as they were.

Appendix 6.1

Marking of Writing took place in January 1999 on the basis of the following four assessment scales:

TASK ACHIEVEMENT

CoE		Description
Operational Proficiency	9	Difficult task completed above required level. (4p)
Vantage	8	Difficult task completed at required level. (3p)
Vantage	7	Difficult task attempted, largely appropriate. (2p) Difficult task attempted, partly appropriate. (1p)
Threshold	6	Difficult task completed below required level.
Threshold	5	Moderate task completed above required level. (4p) Moderate task completed at required level. (3p)
Waystage	4	Moderate task attempted, largely appropriate. (2p) Moderate task attempted, partly appropriate. (1p)
Waystage	3	Moderate task completed below required level.
Breakthrough	2	Simple task above required level. (4p) Simple task at required level. (3p)
Breakthrough	1	Simple task attempted, largely appropriate. (2p) Simple task attempted, partly appropriate. (1p)
Fail	0	Task not attempted; task attempted but completely inappropriate. Handwriting illegible.

GRAMMAR AND SPELLING

CoE		Decription
Operational Proficiency	9	Uses a wide range of complex structures with high accuracy to express ideas. Accurate spelling, though a few slips of the pen might occur.
Vantage	8	Uses a wide range of complex structures with some accuracy to express ideas. Spelling is accurate enough to be followed easily.
Vantage	7	Total control of simple structures. Uses a variety of complex structures, but minor errors may occur. Spelling is accurate enough to be followed easily.
Threshold	6	Good control of a wide range of simple structures. Uses complex structures but errors may occur. Spelling is accurate enough to be followed easily.
Threshold	5	Generally good control of a wide range of simple structures. Uses complex structures with errors but they do not lead to misunderstanding. Spelling is accurate enough to be followed easily.
Waystage	4	Generally controlled use of grammar in a range of simple structures. Grammatical mistakes do not impede comprehension. Some minor spelling mistakes may occur.
Waystage	3	Reasonable control of grammar in a range of simple structures. Grammatical mistakes do not hinder comprehension. A few serious and a number of minor spelling mistakes may occur.
Breakthrough	2	Some control of grammar in restricted range of simple structures. Grammatical mistakes may hinder comprehension. A few serious and a number of minor spelling mistakes may occur.
Breakthrough	1	Little control of grammar in restricted range of simple structures, which occasionally hinders comprehension. Both minor and serious spelling mistakes may occur frequently.
Fail	0	Total lack of control even in simple structures impede comprehension. Full of spelling mistakes, even in basic vocabulary.

VOCABULARY (accuracy and range)

CoE		Description
Operational Proficiency	9	Good command of a broad lexical repertoire including idiomatic expressions and colloquialisms Few avoidance strategies, occasional minor slips but no significant errors
Vantage	8	Good range of vocabulary to express fine shades of meanings both in concrete and more abstract topics Lexical gaps do not hinder comprehension.
Vantage	7	Fair range of vocabulary to express shades of meanings both in concrete and abstract topics Incorrect word choice does not hinder comprehension.
Threshold	6	Appropriate range of vocabulary sufficiently varied in familiar areas Inaccuracies when expressing more complex thoughts
Threshold	5	Sufficient range of vocabulary in familiar topics Regular inaccuracies
Waystage	4	Sufficient (though limited) vocabulary for routine situation, everyday familiar topics with many inaccuracies
Waystage	3	Basic vocabulary repertoire, sufficient only for coping with simple survival needs in familiar topics Inaccuracies may hinder comprehension.
Breakthrough	2	Narrow vocabulary repertoire (sometimes isolated words and phrases) related to everyday topics
Breakthrough	1	Mostly inappropriate limited repertoire (string of words and phrases) Difficult to comprehend
Fail	0	Completely inappropriate

ORGANISATION

CoE		Description
Operational Proficiency	9	expressing ideas and thoughts at length with good arguments and appropriate and varied supporting details fully appropriate layout
Vantage	8	systematically developed topic with good arguments and some supporting details fully appropriate layout
	7	systematically developed topic with varied arguments but few supporting details fully appropriate layout
Threshold	6	mostly clear development of topic with more varied arguments fully appropriate layout
	5	attempts at developing topic with simple arguments appropriate paragraph structure
Waystage	4	text consisting of sentences linked with only a couple of cohesive devices some attempt at structuring text but little or inappropriate use of linking devices
	3	text consisting of occasionally linked sentences some attempt at structuring text but little or inappropriate use of linking devices
Breakthrough	2	full, simple sentences cohesion between sentences no attempt at structuring text according to stylistic features of task
	1	list of lexical items and phrases, no variety of structures no cohesion between structures no attempt at structuring text according to stylistic features of task
	0	no attempt at organising any kind of text, some lexical items or phrases are written, but without any coherence

Appendix 6.2: Scales for Assessing Speaking

OVERALL COMMUNICATION

CoE		Description
Operational Proficiency	9	communicates successfully, can express himself/herself spontaneously, almost effortlessly, is able to organise extended discourse; uses a wide variety of interactive and communicative strategies
Vantage	8	communicates quite fluently and spontaneously with little sign of having to restrict what s/he wants to say; uses a good variety of interactive and communicative strategies
Vantage	7	communication is successful; uses limited variety of interactive and communicative strategies but contributes effectively throughout the interaction
Threshold	6	communication is usually successful; contributes effectively throughout the interaction and rarely requires prompting
Threshold	5	communication generally successful with some prompting; can initiate, maintain and close simple conversation
Waystage	4	communicates successfully in simple conversational exchanges and is mostly able to keep the conversation going; may use avoidance strategies
Waystage	3	communicates quite successfully in simple and direct exchange of information but is not yet able to keep the conversation going; may occasionally hesitate and resort to avoidance strategies
Breakthrough	2	interacts in a simple way with occasional hesitation; is able to indicate when s/he is following or not; communication is only sometimes successful; utterances are mostly in telegraphic style
Breakthrough	1	very hesitant, but some communication takes place; often requires prompting; rarely initiates; can only respond to simple questions; utterances may be fragmented
Fail	0	no/minimal communication which mainly consists of simple repetition; completely dependant on prompts; fails to comprehend questions

GRAMMAR (range and accuracy)

CoE		Description
Operational Proficiency	9	Uses a wide range of complex structures with high accuracy to express ideas.
Vantage	8	Uses a wide range of complex structures with some accuracy to express ideas.
Vantage	7	Total control of simple structures. Uses a variety of complex structures, but minor errors may occur.
Threshold	6	Good control of a wide range of simple structures. Uses complex structures but errors may occur.
Threshold	5	Generally good control of a wide range of simple structures. Uses complex structures with errors but they do not lead to misunderstanding.
Waystage	4	Generally controlled use of grammar in a range of simple structures. Grammatical mistakes do not impede comprehension.
Waystage	3	Reasonable control of grammar in a range of simple structures. Grammatical mistakes do not hinder comprehension.
Breakthrough	2	Some control of grammar in restricted range of simple structures. Grammatical mistakes may hinder comprehension.
Breakthrough	1	Little control of grammar in restricted range of simple structures, which occasionally hinders comprehension.
Fail	0	Total lack of control even in simple structures impede comprehension.

VOCABULARY (accuracy and range)

CoE		Description
Operational Proficiency	9	Good command of a broad lexical repertoire including idiomatic expressions and colloquialisms Few avoidance strategies, occasional minor slips but no significant errors
Vantage	8	Good range of vocabulary to express fine shades of meanings both in concrete and more abstract topics Lexical gaps do not hinder communication.
Vantage	7	Fair range of vocabulary to express meanings both in concrete and abstract topics Incorrect word choice does not hinder communication.
Threshold	6	Appropriate range of vocabulary sufficiently varied in familiar areas Reasonable accuracy in familiar contexts, inaccuracies when expressing more complex thoughts Communication of the essential message is not prevented.
Threshold	5	Sufficient range of vocabulary in familiar topics Regular inaccuracies do not hinder communication but there are many repetitions and paraphrases
Waystage	4	Sufficient (though limited) vocabulary for routine, everyday transactions, familiar situations and topics Inaccuracies and misunderstandings in non-routine situations
Waystage	3	Basic vocabulary repertoire, sufficient only for coping with simple survival needs in familiar topics Inaccuracies may hinder communication.
Breakthrough	2	Narrow vocabulary repertoire related to everyday topics Repetition and frequent basic errors which impede communication Short utterances
Breakthrough	1	Mostly inappropriate limited repertoire (string of words and phrases) Difficult to comprehend
Fail	0	Insufficient vocabulary to convey message (even after prompting by the interlocutor)

SPEECH QUALITY

CoE		Description
Operational Proficiency	9	high level of pronunciation, stress and intonation no significant mother-tongue interference few minor errors
Vantage	8	high level of pronunciation, stress and intonation no significant mother-tongue interference several minor errors
	7	student's output fully comprehensible some mother-tongue interference good level of pronunciation at times
Threshold	6	student's output fully comprehensible some mother-tongue interference
	5	
Waystage	4	student's output mostly comprehensible strong mother-tongue interference
	3	
Breakthrough	2	student's output sometimes incomprehensible very strong mother-tongue interference
	1	
	0	student's output incomprehensible not enough language produced to assess

Appendix 6.3

Standardised Assessment Sheet for Speaking

Candidate's ID number:		**Assessor's ID number:**	
Criterion	**TASK 1**	**TASK 2**	**TASK 3**
OVERALL COMMUNICATION			
GRAMMAR (range and accuracy)			
VOCABULARY (range and accuracy)			
SPEECH QUALITY			