Chapter 10

The Writing Test

Szabó Kinga, Gróf Szilvia and Charles Alderson

In this chapter we describe the Writing tasks that were pilotted in December 1998, discussing what they were intended to test, and the results of the pilotting. It is important to remember that these were the first Writing tasks we had ever trialled and that although in many cases we were aware of possible problems with the tasks, it was felt important to see how things worked – or sometimes did not work – to prove the point. This phase was a key stage in the process of experimentation and learning, in which participants always had to keep one thing in mind: despite their best efforts and intentions they might be working for the wastepaper basket.

In what follows each task will be described in some detail. For each task, the intended level, the task itself and the type of prompt will be given as characterised in the relevant documents – Specifications and Guidelines for Item Writers – as well as a short task description and brief comments on the skills being tested. There then follows an account of the main comments made by markers on the marking criteria and descriptors in the rating scales, and proposals for changes to the rating scales for the future. But first a summary of the various tasks:

Table 10.1: The pilotted Writing tasks

Task ID	Task-type	Type of output text	Verbal/visual prompt(s)	Evaluation
Booklet 1 Task 1 **Intended level: Basic**	form filling	• only words plus ticking	Advertisement from a leaflet – used for creating a situation but not really necessary for completing the task	• realistic task • communicative • standardising answers is not possible • does it test writing or reading??
Booklet 1 Task 2 **Intended level: Basic**	creating a text on the basis of a verbal prompt, a given situation and requirements	• formal letter • about 70 words • transactional	Advertisement from a newspaper/magazine – used for creating a situation but not really necessary for completing the task	• realistic task • communicative • answers will be standardised • some possibility of copying the rubric
Booklet 2 Task 1 **Intended level: Intermed**	creating a text on the basis of a given situation and a verbal prompt	• informal letter • about 100 words • transactional	Advertisement from a leaflet – it contains the requirements for what to write about in verbal form	• realistic task • communicative • standardising answers is not possible • no possibility of copying
Booklet 2 Task 2 **Intended level: Advanced**	creating a text on the basis of a verbal prompt, a given situation and requirements	• formal letter • about 200 words • transactional	Leaflet (advertisement) – used for creating a situation but not really necessary for completing the task	• realistic task • communicative • standardising answers is partly possible, but where it is, there is a possibility of copying much of the rubric

Booklet 3 Task 1 **Intended level:** **Basic**	creating a text on the basis of a given situation and visual prompts	• informal letter • about 80 words • transactional	8 pictures – used for guiding writing so necessary for completing the task	• realistic task • communicative • answers will be standardised • no possibility of copying
Booklet 3 Task 2 **Intended level:** **Basic**	creating a text on the basis of a given situation and a verbal prompt	• informal letter • 80-100 words • transactional	Diary entries – used for guiding writing so necessary for completing the task	• realistic task • communicative • answers will be standardised • little possibility of copying (only words) – acceptable
Booklet 4 Task 1 **Intended level:** **Basic**	creating a text on the basis of a visual prompt, a given situation and requirements	• unclear (letter? report? only sentences as the example suggests?) • about 100 words • transactional	One picture (a bit unclear) – used for guiding writing so necessary for completing the task	• unrealistic task (unfamiliar role: private detective) • answers will be standardised • some imagination required • no possibility of copying
Booklet 4 Task 2 **Intended level:** **Basic**	creating a text on the basis of a given situation and requirements	• letter (formal? informal?) • not longer than 100 words • transactional	None	• realistic task • communicative • standardising answers is not possible • no possibility of copying

Test Booklet 1

Task 1
Intended level: Basic
Task type: form-filling
Prompt: a form
Task description & requirements: filling in data in a form

☞ **This is an advertisement on a leaflet. You would like to be a member of the club, so fill in the form carefully. There are altogether 12 slots and they are numbered (e.g. 4) for you.**

Make new friends in other countries.
Join **Royal Mail's International** Pen Pal Club.
How to become a member?
Fill in the form below and send it back to: Royal Mail, Department 3001,
Freepost NEA 3981, Hendon Road, Sunderland, SR9 9YE, UK.

About you.

Surname _____ (1)

First name _____ (2)

Address _____

_____ (3) Postcode _____ (4)

Date of birth ___ ___ ___ (5)

I write and send _____ (6) letters to _____ (7) friends and/or relatives in _____ (8) countries in a year.

About your Pen Pal.

What country would you like your Pan Pal to come from? Please list three countries.

_____ (9) or _____ (10) or _____ (11)

How old would you like your Pen Pal to be?

Between ____ and ____ years. (12)

Would you like a boy __ or a girl __ ? (please tick)

Send this form back to us and you will make friends in far-away places.

When this item was submitted and accepted, the Specifications for the Basic Writing Examination still included form-filling as an acceptable task type. Nevertheless, it was a source of never-ending debate, and it was hoped that empirical results would support the arguments either way.

Candidates did in fact produce all the problematic answers expected – and even more. Markers had to face numerous problems during the marking procedure: what if data are obviously fake / imaginary? how to assess name / date / address given in the mother tongue instead of the target language and other problems. This was the first time this task type had been tested – and the experience undoubtedly decided its future. Form filling is no longer required on any of the three levels.
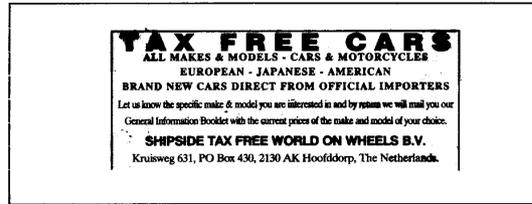
Task 2
Intended level: Basic
Task type: writing a guided formal letter
Prompt: advertisement (approx. 60 words)
Task description & requirements:

This advert appeared in *The European.*

> **TAX FREE CARS**
> **ALL MAKES & MODELS - CARS & MOTORCYCLES**
> **EUROPEAN - JAPANESE - AMERICAN**
> **BRAND NEW CARS DIRECT FROM OFFICIAL IMPORTERS**
> Let us know the specific make & model you are interested in and by return we will mail you our
> General Information Booklet with the current prices of the make and model of your choice.
>
> **SHIPSIDE TAX FREE WORLD ON WHEELS B.V.**
> Kruisweg 631, PO Box 430, 2130 AK Hoofddorp, The Netherlands.

You would like to get a copy of the booklet for your father. Write a letter to the company and

* tell them where you read the ad

* ask them about the Ford Bronco and another model

* ask for a copy

* ask when they will send the booklet

* include a mailing address

Your letter must be about 70 words. Write it on the next page.

The candidates' task was to write a formal letter based on five guiding points, which required candidates to give information, inquire and give personal data (address). They were supposed to produce a letter of about 70 words, using complex sentences, appropriate layout and register.

Test Booklet 2

Task 1
Intended level: Intermediate
Task type: writing a guided informal letter
Prompt: leaflet
Task description & requirements:

This is an advertisement on a leaflet. You would like to be a member of the club, so write your first letter according to the instructions in about one hundred words.

> ## Make new friends in other countries. Join Royal Mail's International Pen Pal Club.
>
> How to become a member?
>
> Fill in the form and write a short letter about yourself. Send them back to: Royal Mail, Department 3001, Freepost NEA 3981, Hendon Road, Sunderland, SR9 9YE, UK.
>
> Letter writing is simple – and fun!
>
> Letter writing is easier than you think. Your letters don't have to be long – in fact, it's best to keep them simple, at first. Start by telling you new Pen Pal all about yourself, your family and your life; what you do at school and the weekends, your favourite food, music, sport, hobby etc. Then round your letter off by asking what life is like in your Pen Pal's country. Once you've got to know each other, finding things to write about will be no problem.
>
> **Send your letter back to us and you will make friends in far-away places.**

Candidates had to write a letter of introduction of about 100 words, following guiding points that were given in a Pen Pal Club's leaflet. In addition to the guiding points, some useful tips were given regarding structure and style. Candidates had to read over 100 words, but the language was quite simple and easy to follow for anyone at the expected level. As most of the input text was a list of suggested points to write about, candidates could lift some useful vocabulary items but it was believed that they had to add their own efforts to shape the text to reflect the best of their English. All these topics must have been familiar to candidates and they were certainly able to relate to each.

Task 2
Intended level: Advanced
Task type: writing a guided formal letter
Prompt: leaflet / advertisement
Task description & requirements:

This is a leaflet of an agency offering you exciting opportunities in the USA. Write a letter to find out more. Make sure you include the following: what you need to do, how soon you can leave, how much and what type of work experience you need, if you can choose where and how long you would like to stay. Give a short description of your work experience and enquire about any other questions you find appropriate after reading the advertisement. You letter should be about 200 words long.

Discover the USA!

If you are under 26 years of age apply now!

√ Free preparation course in USA
√ Free return flights
√ Free medical insurance
√ Departures every week
√ Bonus travel month
√ Official 12 month visa
√ $ 139 a week
√ $ 500 for studies
√ Two weeks paid vacation
√ 24 hour year-round support service
√ Discounted flights for visitors

Interested? Contact
   EF Agency
   5 Kensington Church St, London W8 4LD
   Freephone 0800 163 786
   Internet: http//www.ef.com

Candidates had to describe relevant background experience they were assumed to have and also to inquire about several matters in order to 'discover the USA'. But as a starting point they had to read the task – which together with the rubrics constituted a text of over 130 words. The guiding points were all given in the rubric and the exact number of points to be covered was left fairly open as candidates were instructed to ask 'any other questions' they might find appropriate after reading the advertisement. Candidates had to find out about details of the possible job and write about their work experience in about 100 words.

Test Booklet 3

Task 1
Intended level: Basic
Task type: writing a guided informal letter
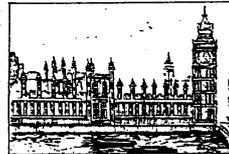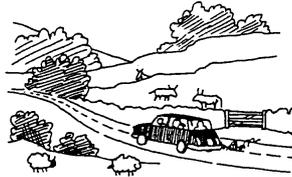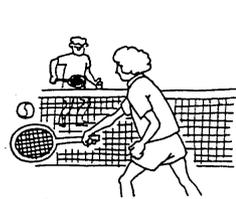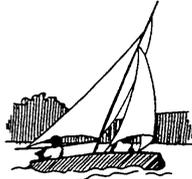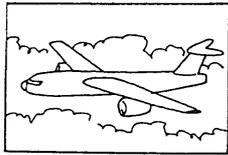**Prompt:** visual prompts (drawings)
Task description & requirements:

👉 Your English pen-friend, Michael, has invited you to England this summer. Now you are staying with his family. They live about 50 miles far from London.

Write a letter home to your parents about the first 10 days of your visit. Use the pictures below.

Write your letter on this page, using about 80 words.

Candidates had to use pictures as a guide to write a letter of about 80 words. This task did not require a lot of reading as only the instructions were given in English, and the language is hopefully simple enough for a 'Basic' candidate to understand. Candidates were expected to use narrative and descriptive forms.

Task 2
Intended level: Basic
Task type: writing a guided informal letter
Prompt: diary entries
Task description & requirements:

☞ **Write a letter to your friend about your last week based on the diary entries below.**

Write 80-100 words.

Monday	*dentist*
Tuesday	*no school ! touring with friends* *- bike goes wrong on way home*
Wednesday	*no homework - mark 1*
Thursday	*see Titanic (3rd time)* *Pizza with friends*
Friday	*neighbours move*
Saturday	*visit grandma with family*
Sunday	*letter!*

Candidates were required to use notes entered in a weekly diary and write a letter to a friend telling about the week in 80-100 words. There were several different activities and happenings they were expected to include in the text, all of which were likely to have happened to any Hungarian secondary student, so it should not have been difficult for them to relate to the task.

Test Booklet 4

Task 1
**Intended level:** Basic
Task type: writing a description
Prompt: picture (drawing)
Task description & requirements:

Task 1. 'The Private Eye'
You are a private detective. Your client asked you to follow a certain Edith, who works in this office. On the basis of what you can see in her office, write a short description of Edith in about 100 words. Think of her job / daily routine / clothes / family life / hobby / favourite food etc.

Example: I think she can operate a computer.
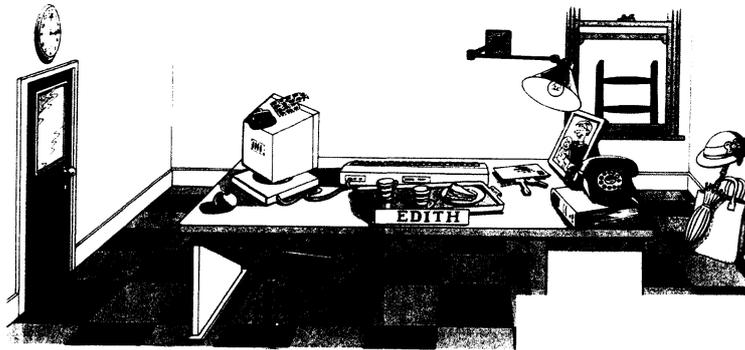
Candidates had to use a drawing of an office in order to arrive at conclusions concerning the person who works in this office. They were given suggestions as to what aspects they should identify. The drawing was rich in prompts so if candidates only gave a detailed description of the room they would not have completed the task which was to write 100 words about the person.

Task 2
Intended level: Basic
Task type: writing a letter of introduction
Prompt: none
Task description & requirements:

Task Two

You have a chance to go on a two-week language course in England. You will stay with a family there. Your host family would like to get some information about you before you arrive. Write a short letter to them in which you write about the following points:
- your family
- your home
- your school
- your favourite free time activities; likes/dislikes; pets; favourite meals; etc.
You do not need to write any address. Your letter must not be longer than 100 words.

Write your letter here:

The guiding points were given in the rubric for the task. Candidates had to write a letter to a future host family they were to stay with in order to introduce themselves. They had to write about familiar topics and aspects of their life in up to 100 words.

Overall comments from markers on the Writing tasks

Task-types

All but one of the task-types were similar to the ones which are likely to be included in the Writing Paper of the School-leaving examination. That is, candidates will be required to create texts on the basis of a given situation and given requirements and/or visual/verbal prompt(s). The only exception was the form-filling task (1/1), which will not be included in the examination for several reasons: (1) it is debatable whether it tests reading or writing; (2) in most cases it is not possible to standardise items and so marking will be difficult, as we have seen; (3) in our setting it would be difficult to gather authentic forms the topic of which is appropriate for the tested population and which would contain enough items.

Another similarity to the envisaged future examination was that each piloted task attempted to create a communicative writing task, in which the situation in which to write, the reader, the purpose and the topic of the output text were made clear to the candidate as well as the relationship between the reader and the writer. All these determined the text-type, the style and the register according to which candidates had to create their texts.

3/1 was not a realistic task as candidates were asked to write a letter back home to their parents telling about their stay abroad in English. At the same time it is interesting to see that when completing 4/2 where the task was to write a letter of introduction to a future host family, some candidates misunderstood the task and wrote a letter home to their parents telling about their adventures in English.

It should be noted that we piloted very few tasks at levels other than Basic: only one was targetted at the Intermediate level, and one at Advanced. We will discuss the empirical results below, but it is clear from Table 10.5 that the so-called Basic tasks varied in difficulty because of differences in markers rather than task design, and in fact candidates achieved at least as high mean scores on the Intermediate and Advanced tasks as they did on Basic tasks. The issue of level of difficulty of task cannot (yet) be separated from that of the severity of the markers.

Types of output texts

Markers commented on the fact that there were two letter writing tasks in Booklets 2 and 3, which did not give candidates the opportunity to show their competence in other genres. It was recommended that there should only be one letter writing task in the final exam.

Six out of the eight tasks required students to write a letter, 3 of which were meant to be informal, 2 formal, but in one task (4/2: a letter to an unknown future host family) the required degree of formality was less clear, which might have caused a problem to students as well as to markers. One task was a form-filling task (1/1), which required students only to write words and a tick, in which case the answers could not be claimed to be a text. In one task (4/1) it was unclear what text-type, if any, was required, and since an example sentence was given, the task seemed to suggest that writing individual sentences rather than a cohesive text was required.

Whereas writing letters, both formal and informal, will be a possible (and important) task in the Writing Paper – and it must be acknowledged that it is the most obvious

communicative writing task, especially if we want to avoid traditional composition-writing tasks – , it must not be forgotten that other text-types are planned to be included, such as messages, postcards and greeting cards, reader's letters, articles for student magazines, applications, CVs and their cover letters. In order to experiment with these text types it is desirable that they be included in the next pilot. These text-types would make it possible to elicit more interactional language as well as argumentative texts, whereas the piloted tasks focussed on transactional language only.

As to the length and nature of the output texts, our revised plans are to require students to write one shorter transactional/interactional text (Part 1), and one longer argumentative text (Part 2). Thus Task 2/2 does not fit this framework. Also, decisions will have to be made on the level of tasks since the tasks for the Advanced level examination will require slightly longer output texts than the ones for the Intermediate level examination.

Rubrics and prompts

When examining the rubrics and the prompts of the piloted tasks, two main questions have to be considered: How and to what extent do they determine what students should write? To what extent do they provide the language to be used (i.e. is it possible to copy them)?

Four of the Writing tasks used an advertisement as prompts, two of them used a picture or pictures, one used diary entries, and in one case there was no prompt given. Whereas these prompt-types are perfectly acceptable and are in agreement with the Specifications, again it must be noted that a greater variety of both verbal and visual prompts, including notes, messages, short articles, letters, graphs, charts, tables, picture sequences, cartoons, etc., would be desirable. The fact that Task 4/2 used no prompts does not in itself make the task unacceptable, as this task successfully created a communicative task by giving a situation and listing the necessary content in bullet points; however, this task is problematic for other reasons pointed out below.

A more careful examination of the prompts reveals that some of them are not real prompts in the sense that they do not actually define what students have to write about: see Tasks 1/1, 1/2 and 2/2. It is worth considering to what extent, if any, these prompts are actually needed and whether they help or confuse/mislead/distract candidates.

The advertisement used in 2/1 is a more successful prompt in the sense that it gives guidance (indeed instructions) to students in terms of what to write about. Similarly, the task using diary entries (3/2) is equally successful in guiding candidates, and although some of the language to be used is thus provided, it is given in note form, therefore candidates still have to use all their abilities to create continuous prose.

The two tasks with visual prompts (Task 3/1 with eight pictures, Task 4/1 with one picture) are perhaps the most successful ones in terms of determining the content of the output text without providing any of the language to be used, and at the same time ensuring standardised answers. This issue will be discussed in more detail below.

When the requirements as to the content of the output text are listed in the rubric, it is worth examining how these are formulated and organised. Task 1/2 gives them in clear, easily understandable language, and organises them in bullet points so they are easy for students to follow and the danger that they might leave out something is minimal. The rubric in Task 4/2 shows the same attempt, but the language level of the task is probably too high for this very elementary task. The rubric in Task 2/2 is less successful since the requirements are given in continuous text, and most of them are such (reported questions) that by pure sentence transformation they can be turned into the required sentences. The final requirement ('enquire about any other questions you find appropriate after reading the advertisement') is problematic, as it is not clear how many more questions, if any,

candidates have to write, and thus task achievement will be very difficult to define and therefore to mark.

Whether the content of the output text is determined by prompts or the rubric, it is very important to emphasise to candidates that they are required to write about **all** the requirements (e.g. about all the pictures, all the entries in a diary, all the questions given, etc.) Unless we do so, we cannot fairly evaluate task achievement. It must also be noted that 'etc.' in rubrics is highly problematic for the same reason (Tasks 2/1, 4/1 and 4/2 include this in their requirements).

The issue of standardising answers

To what extent answers in Writing tasks can and should be standardised is still a question open to debate. One might argue that it is not realistic, and perhaps not fair, to require candidates to write about events that have never happened to them, as is the case in Tasks 3/1 and 3/2. On the other hand, tasks which only determine **what** students have to write **about**, but not **what** they have to write might encourage them to use avoidance strategies (i.e. they might write not what is true about them but what they can write in English) as well as to write down rote-learnt material, as is the case in Tasks 1/1, 2/1, 4/2, and part of Task 2/2, where candidates are required to write about themselves.

Marking Scales and Criteria

The original marking scales were presented in Chapter 6, Appendix 6.1.

Markers were asked to comment on their experience of marking and on each criterion.

- Booklet 1 Task1 was supposed to be an objectively assessed task. Unfortunately, markers did not agree how to assess the task before marking, so there are likely to have been great differences in what and how they actually assessed.
- One marker found it extremely difficult to ignore grammatical mistakes when it came to marking other criteria. She felt that grammatical weaknesses caused her to underline candidate's performance in other aspects, too.
- The problem of length was not addressed before marking began and many markers asked questions like: If the script is much longer than required how should that be treated? If it is much longer can it be at the required level? What is the criterion for an achievement of 50% only? Who is to blame if the candidate wrote twice as much as or only half the required length – maybe only the minimum length should be given in the task? If the candidate sticks to the 80-100 word limit does that not restrict him/ her? What can or cannot we find out about his command of the language?
- Many more sample performances are needed to standardise markers' understanding of what can be expected on these tasks.
- There are probably too many levels on any one scale.

TASK ACHIEVEMENT
- As markers were not told which level a given task was written for, it was felt quite likely that depending on their personal interpretation there would be quite a spread of Task Achievement scores. Almost every marker complained that in the situation where they did not know the intended level, they found it impossible to use the scale for Task Achievement. On the other hand the only positive feedback also came in connection with Task Achievement: markers liked the idea that an element of differentiation was involved in each band, that is they were instructed to recognise top and bottom performances. It was recommended that the idea should be adopted for the other criteria.
- If a candidate misunderstood or altered a task (or a part of it) what should happen? Presumably the candidate scores 0 for Task Achievement but what about the rest? (One

candidate changed or misunderstood the task. Instead of writing in the Past Tense s/he used the Future. The text itself is not bad but the task has not been fulfilled.)

GRAMMAR AND SPELLING
- It depends on the design of the task as to whether the candidate uses 'a wide range of complex structures' – even if s/he is able to, the task must be such that it gives scope for the use of complex structures. 'Uses complex structures...' in bands 5 and 6 should be replaced with 'If uses complex structures...' thus overcoming the problem.
- Spelling is not so important. It should have less weight.

VOCABULARY
- Some aspects are highly task dependent: if the task does not give scope for 'fine shades of meaning', however good the performance may be, it cannot be rated above band 6.
- '...when expressing more complex thoughts'. As the task appeared simple, even more able candidates may not have felt the need to express complex thoughts.
- There were no 'abstract topics'. Does that mean that no candidate can / should score higher than band 6?
- Must there be regular inaccuracies (band 5) – does it mean a candidate's performance is band 6 if s/he commits no regular inaccuracy? Can candidates commit many inaccuracies if they are only required to write a short piece?
- Do above required level and required level belong to the same band? Do largely appropriate and partly appropriate belong to the same band?
- 'Familiar' topics is vague – different things may be familiar for different candidates.
- No samples are given for the top 3 bands – they don't seem to be built on one another.

ORGANISATION
- This depends on the task – what was the expected layout for the detective-task? Can candidates only score 0 or 1 here?
- What if layout is appropriate (e.g. the piece is a letter) but the text is not structured into paragraphs? If the text is clear, easy to follow, logically built, yet there is no 'visual' segmentation? Can cohesion be 'unexpressed'?
- Why is 'cohesion between the sentences' introduced at level 2 and not mentioned at higher levels?
- 'Proper use of paragraphs' is introduced in band 5 – what if the letter is very logically built but indentation is missing at the beginning of each new section?
- Why is 'according to stylistic features of task'' only mentioned at levels 1 and 2?

Empirical results from the pilotting

In Chapter 8, we presented an analysis of the results of the Writing tasks and the reliability of marking. Here we go into more details on how the marking criteria were used, and how reliable the marking was.

We saw in Chapter 8 that the overall reliability coefficients were very respectable, but the data conflated the tasks and rating criteria. In what follows, we examine the results in more detail. Table 10.2 breaks down results for the different criteria, by first and second task.

Table 10.2: Inter-rater agreement, by criterion, for all booklets except Booklet 1

	Task 1	Task 2
Ach	.38 (n=191)	.55 (n=177)
Grm	.77 (n=189)	.79 (n=173)
Voc	.66 (n=189)	.76 (n=173)
Org	.70 (n=189)	.77 (n=173)

These average correlations for first and second raters show only moderate inter-rater reliability for most criteria, and unacceptably low reliability for Task Achievement, and for Vocabulary on the average Task 1. However, as with the total score, it is important to look at the scores given for each criterion, task by task, to see whether there are problems with particular tasks. Table 10.3 reports these results:

Table 10.3: Inter-rater agreement (rater 1 with rater 2), by criterion, task by task

		Task 1	Task 2
Booklet 1			
	Ach	–	.61 (n=66)
	Grm	–	.81 (n=65)
	Voc	–	.79 (n=65)
	Org	–	.81 (n=65)
Booklet 2			
	Ach	.39 (n=67)	.66 (n=55)
	Grm	.83 (n=67)	.79 (n=52)
	Voc	.67 (n=67)	.80 (n=52)
	Org	.72 (n=67)	.83 (n=52)
Booklet 3			
	Ach	.33 (n=63)	.40 (n=60)
	Grm	.83 (n=62)	.85 (n=59)
	Voc	.71 (n=62)	.76 (n=59)
	Org	.78 (n=62)	.74 (n=59)
Booklet 4			
	Ach	.36 (n=61)	.38 (n=62)
	Grm	.63 (n=60)	.77 (n=62)
	Voc	.61 (n=60)	.67 (n=62)
	Org	.63 (n=60)	.69 (n=62)

These results are the most interesting and relevant to the revision of tasks and the inspection of criteria, since they present results task by task, and represent large enough numbers of assessments per task to give confidence in the reliability of results.

They confirm the problem with Task Achievement on Task 1 in particular, although Task 2 in Booklets 3 and 4 are also problematic in relation to this criterion. The only figure approaching respectability is Task Achievement on Task 2, Booklet 2, and this task needs to be inspected and compared with the other tasks (see above discussion).

Vocabulary seems to be a problem on Booklet 4, both tasks, and Booklet 2, Task 1. Again, a comparison of these tasks should reveal greater insight into tasks and this particular criterion.

Organisation is problematic in Booklet 4, both tasks, whereas reasonable reliability is achieved on Task 2 in Booklets 1 and 2.

Grammar and Spelling is the criterion where most agreement seems to be reached, perhaps not surprisingly in the Hungarian context, but even here there is worrisome lack of agreement in Booklet 4, especially Task 1.

We must, however, add the caveat that we are not comparing individual raters in Table 10.3, but are aggregating raters, since 'rater 1' includes raters 1, 2, 3, 4, 5, 6 and 8, and 'rater 2' includes raters 4, 5, 7, 9, 10, 11 and 14.

Table 10.4 below compares the performance of pairs of individual raters, criterion by criterion, for aggregations of Task 1, and also Task 2.

Table 10.4: Inter-correlations of criterion scores given by pairs of raters

Pairs:		Task 1	Task 2
5:10	Ach	.95 (n=18)	.74 (n=23)
	Grm	.86 (n=18)	.88 (n=22)
	Voc	.86 (n=18)	.84 (n=22)
	Org	.77 (n=18)	.80 (n=22)
3:5	Ach	.50 (n=18)	.85 (n=23)
	Grm	.71 (n=18)	.82 (n=23)
	Voc	.69 (n=18)	.93 (n=23)
	Org	.76 (n=18)	.82 (n=23)
1:14	Ach	NS (n=11)	.72 (n=13)
	Grm	.94 (n=10)	.83 (n=11)
	Voc	.82 (n=10)	.70 (n=11)
	Org	.86 (n=10)	NS (n=11)
2:7	Ach	NS (n=36)	.47 (n=48)
	Grm	.77 (n=36)	.76 (n=48)
	Voc	.68 (n=36)	.82 (n=48)
	Org	.70 (n=36)	.86 (n=48)
6:9	Ach	.76 (n=34)	.65 (n=46)
	Grm	.85 (n=34)	.87 (n=45)
	Voc	.37 (n=34)	.57 (n=45)
	Org	.84 (n=34)	.84 (n=45)
8:11	Ach	.56 (n=34)	.74 (n=39)
	Grm	.74 (n=34)	.83 (n=39)
	Voc	.70 (n=34)	.88 (n=39)
	Org	.75 (n=34)	.81 (n=39)
3:10	Ach	.56 (n=16)	NS (n=20)
	Grm	.87 (n=15)	.73 (n=18)
	Voc	.87 (n=15)	.80 (n=18)
	Org	.74 (n=15)	.71 (n=18)
1:4	Ach	NS (n=16)	.82 (n=15)
	Grm	.84 (n=16)	.78 (n=15)
	Voc	.66 (n=16)	.82 (n=15)
	Org	.62 (n=16)	.80 (n=15)

What this table shows is quite high agreement amongst the individual criteria for most pairs of raters, although pairs vary in their agreement. However, Task Achievement again clearly causes problems – often resulting in no correlation between raters, especially on Task 1. Vocabulary again also seems to cause problems for some raters, with unacceptably low coefficients.

As we saw in Chapter 8, it is important not only to look at correlations, but also at mean scores. In Chapter 8, we saw that raters do indeed differ from each other in the harshness or leniency with which they rated scripts.

As we see in Table 10.5 below, the difficulty of tasks varied, depending on who rated the task:

Table 10.5 Differences in mean scores for first and second raters, by task

| | | Rater 1 | | Rater 2 | | | | |
		Mean	sd	Mean	sd	n	t	p
Booklet 1								
	Task 2	13.62	7.15	11.981	6.57	65	3.92	.000
Booklet 2								
	Task 1	15.79	5.785	13.85	5.847	67	4.78	.000
	Task 2	14.10	9.725	12.12	9.510	50	3.01	.004
Booklet 3								
	Task 1	13.63	6.178	13.42	7.436	62	0.34	.735
	Task 2	13.80	6.411	12.59	7.247	59	2.29	.026
Booklet 4								
	Task 1	13.77	5.030	12.32	5.193	60	2.98	.004
	Task 2	14.37	5.038	12.11	5.726	62	4.65	.000

+ n = number of scripts; t = t. value; p = probability

Thus, means are significantly different between first and second raters for all booklets except for Booklet 3 Task 1. It is, however, difficult to generalise and say that, for example, Booklet 1 Task 2 is harder than other tasks, despite the mean for second raters (11.981) since the average first rater gave a mean of 13.62, which is as high as means for several other tasks. Similarly, although Booklet 2 Task 1 got the highest mean, and so appears easiest, this was only the case for the average first rater (15.79). The average second rater gave a mean of 13.85, which is no higher than several other tasks. This shows clearly the interaction between rater and task in the apparent 'difficulty' of a task, and emphasises yet again the importance of rater training before difficulty indices can be believed for any given task.

It must be remembered that we are here dealing with overall scores, and not scores for the different criteria. To see how marking criteria and tasks interacted, we need to compare how raters used the various criteria, by tasks, aggregated across raters.

Table 10.6 below shows the mean scores for each criterion, averaged over all raters, by task.

Table 10.6: Average scores by criterion, by task

Booklet 1						
Task2						
	Variable	Mean	sd	Minimum	Maximum	n
	ACH	2.91	1.72	.00	5.00	134
	GRM	3.00	1.93	.00	9.00	133
	VOC	3.53	1.96	.00	8.00	133
	ORG	3.14	1.90	.00	6.00	133
Booklet 2						
Task 1						
	ACH	3.43	1.44	.00	7.00	134
	GRM	3.50	1.72	.00	8.00	134
	VOC	4.25	1.77	.00	8.00	134
	ORG	3.65	1.74	.00	8.00	134
Task 2						
	ACH	2.87	2.63	.00	9.00	121
	GRM	2.78	2.42	.00	8.00	115
	VOC	3.17	2.63	.00	8.00	115
	ORG	2.94	2.55	.00	8.00	115
Booklet 3						
Task 1						
	ACH	2.87	1.82	.00	7.00	126
	GRM	3.10	2.01	.00	8.00	125
	VOC	3.91	1.98	.00	9.00	125
	ORG	3.51	1.91	.00	8.00	125
Task 2						
	ACH	2.55	1.59	.00	6.00	123
	GRM	3.03	2.11	.00	9.00	121
	VOC	3.80	2.02	.00	8.00	121
	ORG	3.55	1.99	.00	8.00	121
Booklet 4						
Task 1						
	ACH	2.59	1.30	.00	5.00	123
	GRM	3.05	1.63	.00	7.00	122
	VOC	3.91	1.56	.00	7.00	122
	ORG	3.26	1.77	.00	8.00	122
Task 2						
	ACH	2.62	1.49	.00	6.00	124
	GRM	3.23	1.67	.00	7.00	124
	VOC	4.00	1.57	.00	8.00	124
	ORG	3.40	1.57	.00	7.00	124

From Table 10.6, it is clear that Vocabulary usually receives the highest score. This does not necessarily mean that candidates' writing is best from the point of view of vocabulary. Maybe raters are particularly lenient when it comes to judging students' vocabulary knowledge or usage, or these tasks are not capable of challenging a candidate's lexical knowledge.

Task Achievement is typically scored lowest, with Grammar following. This does not necessarily mean that students were weak at understanding and complying with the task set: it is important to remember that Task Achievement was also marked the most inconsistently. That raters marked Grammar and Spelling fairly harshly and also consistently is consistent with the view that Hungarian teachers focus particularly on grammar mistakes and are used to doing so.

In summary, it is clear that many of the problems that emerged in the pilotting of the Writing tasks are at least as much due to the wording of the rating scales as to problems of task design. Although we have identified some problems that will need to be attended to in future rounds of task development, the biggest problem seems to be the need for new rating scales, and the training of markers in their use.

Changes to the Writing Assessment Scales

The first draft of the new marking scales for assessing written performance was prepared in January 2000. The draft attempts to preserve what was felt to be good in the original scales but several changes were made so that hopefully the scales will be easier to use and will reflect to a larger extent what we intend to assess in written performance. See Appendix 10.1 for the new scales.

The most obvious change is that the new scales are level-specific, i.e. separate scales have been prepared for the Intermediate-level and the Advanced-level examinations. This made it possible for us to have fewer bands (8 instead of the original 10), only 5 of which (bands 0, 1, 3, 5 and 7) are defined with descriptors. The intention is that markers will have to read and remember much less, which will hopefully result in greater reliability of marking. Also, level-specific scales make it possible to assess how well candidates perform on tasks of a given level. We acknowledge the fact that tasks of different levels elicit language of differing degrees of difficulty, and place different demands on candidates in terms of task achievement, organisation, etc.

The four main criteria of the original scales have been kept, but with slightly different headings and content. The new criteria are the following: (1) Task achievement & content instead of the original Task achievement, (2) Grammar & spelling, (3) Vocabulary, (4) Organisation, cohesion & layout instead of the original Organisation. With these new headings we intend to lay more emphasis on what is included in each criterion. Where necessary, the number of sub-criteria within a criterion have been increased, and care was taken to ensure that each band describes the same sub-criteria (where appropriate). The sub-criteria are organised in bullet points for greater transparency.

Thus, in (1) Task achievement & content the descriptors describe the extent to which the task is completed, the number of content points covered, the consistency of style and register, and the effect on the reader, whereas the original scales described only whether the task was attempted/completed and how appropriately/at what level. The descriptors for assessing (2) Grammar & spelling and (3) Vocabulary describe the same sub-criteria as the original scales did, but in a more systematic, therefore more usable, way. The same is true of (4) Organisation, cohesion & layout, where the new scales assess organisation of the text, the use of cohesive devices, and paragraphing and layout, all of which were included in some, but not all, of the descriptors of the original scales.

As pointed out above, three of the bands in the new scales do not have descriptors. These have been left blank to allow markers to grade performances which would fall into two different bands according to the different sub-criteria involved. Band 3 is intended to be a kind of 'pass-mark' in the sense that it describes the performance which is adequate for the level in question. As the total marks are 28, a minimum of 12 marks would thus be

needed for a performance to reach the required standard. However, we still need to consider whether a sub-standard performance on one criterion could be compensated by a better performance on other criteria.

The Writing Paper of the new School-leaving examination, as currently envisaged, will include a shorter and a longer task. One issue to be resolved is whether to assess the two tasks together since one task in itself (especially a shorter one) might not be able to elicit all the features of written performance to be assessed. Alternatively, each task could be marked separately, and then the two ratings could be combined in such a way as to allow the more demanding task to carry more weight.

In addition to the new general marking scales, we will need to produce task-specific marking schemes every year to take account of the differing demands of each task. More experience and experimentation with such task-specific scales is needed. In addition, systematic and thorough training of markers will be required and the score bands will need to be illustrated by identifying benchmark scripts from pilotting. Double-marking is essential, and systems for the external monitoring of the quality of the ongoing marking will also need to be developed. All these are essential in order to make the assessment of the Writing Paper more objective and reliable.

Appendix 10.1: Assessment scales for the Writing test – Intermediate level – Version 4 – February 2000

Az írásbeli feladatok értékelési szempontrendszere – Középszint

	Task achievement, content	Grammar, spelling	Vocabulary	Organisation, cohesion, layout
7	In intermediate level tasks • task fully completed with all content points covered in required detail • required register and style used consistently • very positive effect on target reader	In intermediate level tasks • consistently accurate control of grammar and spelling • wide range of structure (where appropriate) • minimal errors only, which do not interfere with understanding	In intermediate level tasks • appropriate and varied vocabulary	In intermediate level tasks • effective organisation • a variety of cohesive devices • fully appropriate paragraphing and layout
6				
5	• task is completed with all content points covered • inconsistencies in register or style may occur • positive effect on target reader	• good control of grammar and spelling with minor inaccuracies • a range of structure (where appropriate) • some errors, which do not impede understanding	• appropriate vocabulary	• clear organisation • suitable cohesive devices • mostly appropriate paragraphing and layout
4				
3	• task attempted but not fully achieved with relevant number of content points covered • would achieve the required effect on target reader	• adequate control of grammar and spelling • adequate range of structure • errors may sometimes hinder understanding	• basic but mostly appropriate vocabulary	• adequate organisation • simple but appropriate cohesive devices • layout and paragraphing is generally appropriate but may show some inconsistencies
2				
1	• inadequate attempt at task with most content points not covered • would have a negative effect on target reader	• inadequate control of grammar and spelling • narrow range of structure • frequent errors occur even in basic structures	• basic and/or inappropriate vocabulary	• inappropriate organisation • very few or mostly inappropriate cohesive devices • inappropriate paragraphing and layout
0	task not attempted OR task attempted but content is inappropriate OR handwriting illegible	total lack of control even in simple structures OR too little language for assessment	completely inappropriate vocabulary OR little language for assessment	total lack of organisation and cohesive devices OR too little language for assessment

	A kommunikatív cél teljesítése, tartalom	Nyelvhelyesség, helyesírás	Szókincs	A szöveg felépítése, szövegkohézió, elrendezés
7	Középszint• feladatokban • a vizsgázó a feladatot teljes egészében megoldotta, minden tartalmi egységr•l a megkívánt részletességgel írt • a vizsgázó következetesen a megfelel• stílust és regisztert alkalmazta • a szöveg nagyon jó benyomást tenne a megcélzott olvasó(k)ra	Középszint• feladatokban • következetesen helyes nyelvtani szerkezetek és helyesírás • széles skálán mozgó, változatos nyelvtani szerkezetek (ahol erre a feladat módot ad) • csak egy-két kisebb hiba fordul el•, melyek a megértést nem befolyásolják	Középszint• feladatokban • odaill•, változatos szóhasználat	Középszint• feladatokban • jól felépített szöveg • a szövegkohéziót biztosító eszközök használata változatos • a szöveg bekezdésekre tagolása és elrendezése teljesen megfelel•
6				
5	• a vizsgázó a feladatot megoldotta, minden tartalmi egységr•l írt • a szöveg stílusa és a regiszter alkalmazása nem teljesen következetes • a szöveg jó benyomást tenne a megcélzott olvasó(k)ra	• helyes nyelvtani szerkezetek és jó helyesírás, kisebb pontatlanságokkal • változatos nyelvtani szerkezetek (ahol erre a feladat módot ad) • néhány hiba el•fordul, de ezek a megértést nem akadályozzák	• odaill• szóhasználat	• világosan felépített szöveg • a szövegkohéziót biztosító eszközök használata megfelel• • a szöveg bekezdésekre tagolása és elrendezése többnyire megfelel•
4				
3	• a vizsgázó a feladatot elkezdte, de nem oldotta meg teljes egészében, az adott feladathoz az Értékelési útmutatóban meghatározott számú tartalmi egységr•l írt • a szöveg megfelel• benyomást tenne a megcélzott olvasó(k)ra	• megfelel• nyelvhelyesség és helyesírás • a nyelvtani szerkezetek még elfogadhatóan széles skálája • a hibák id•nként gátolják a megértést	• egyszer•, de többnyire odaill• szóhasználat	• megfelel•en felépített szöveg • egyszer•, de odaill• szövegkohéziót biztosító eszközök • a szöveg bekezdésekre tagolása és elrendezése általában megfelel•, de következetlenségek el•fordulnak
2				
1	• A vizsgázó nem megfelel•en fogott hozzá a feladat megoldásához, a legtöbb tartalmi egységr•l nem írt • a szöveg rossz benyomást tenne a megcélzott olvasó(k)ra	• nem megfelel• nyelvhelyesség és helyesírás • a nyelvtani szerkezetek sz•k skálája • még a legegyszer•bb szerkezetekben is gyakoriak a hibák	• egyszer• és/vagy oda nem ill• szóhasználat	• nem megfelel•en felépített szöveg • nagyon kevés vagy többnyire nem megfelel• szövegkohéziót biztosító eszközök • a szöveg bekezdésekre tagolása és elrendezése nem megfelel•,
0	a vizsgázó el sem kezdte a feladatot, VAGY hozzákezdett, de a szöveg tartalma nem megfelel•, VAGY olvashatatlan a kézírás	még a legegyszer•bb szerkezetek is teljesen hibásak, VAGY túl kevés a szöveg az értékeléshez	egyáltalán oda nem ill• szóhasználat, VAGY túl kevés a szöveg az értékeléshez	a szöveg egyáltalán nincs felépítve, és nincsenek szövegkohéziót biztosító eszközök, VAGY túl kevés a szöveg az értékeléshez

Advanced level – Version 2 – February 2000

Az írásbeli feladatok értékelési szempontrendszere – Emelt szint

	Task achievement, content	Grammar, spelling	Vocabulary	Organisation, cohesion, layout
7	In advanced level tasks • task fully completed with all content points covered in required detail • required register and style used consistently • very positive effect on target reader	In advanced level tasks • consistently accurate control of grammar and spelling • wide range of structure (where appropriate) • minimal errors only, which do not interfere with understanding	In advanced level tasks • precise and varied vocabulary	In advanced level tasks • effective organisation • a variety of cohesive devices • fully appropriate paragraphing and layout
6				
5	• task is completed with all content points covered • inconsistencies in register or style may occur • positive effect on target reader	• good control of grammar and spelling with minor inaccuracies • a range of structure (where appropriate) • errors in more complex structures do not impede understanding	• varied vocabulary	• clear organisation • appropriate cohesive devices • mostly appropriate paragraphing and layout
4				
3	• task attempted but not fully achieved with relevant number of content points covered • would achieve the required effect on target reader	• adequate control of grammar and spelling • adequate range of structure • a number of errors which do not impede understanding	• appropriate vocabulary	• adequate organisation • appropriate but often simple cohesive devices • layout and paragraphing is generally appropriate but may show some inconsistencies
2				
1	• inadequate attempt at task with most content points not covered • would have a negative effect on target reader	• inadequate control of grammar and spelling • narrow range of structure • frequent errors obscure understanding	• limited and/or inappropriate vocabulary	• inappropriate organisation • very few or mostly inappropriate cohesive devices • inappropriate paragraphing and layout
0	task not attempted OR task attempted but content is inappropriate OR handwriting illegible	total lack of control even in simple structures OR too little language for assessment	completely inappropriate vocabulary OR little language for assessment	• total lack of organisation and cohesive devices OR too little language for assessment

	A kommunikatív cél teljesítése, tartalom	Nyelvhelyesség, helyesírás	Szókincs	A szöveg felépítése, szövegkohézió, elrendezés
7	Emelt szint• feladatokban • a vizsgázó a feladatot teljes egészében megoldotta, minden tartalmi egységr•l a megkívánt részletességgel írt • a vizsgázó következetesen a megfelel•stílust és regisztert alkalmazta • a szöveg nagyon jó benyomást tenne a megcélzott olvasó(k)ra	Emelt szint• feladatokban • következetesen helyes nyelvtani szerkezetek és helyesírás • széles skálán mozgó, változatos nyelvtani szerkezetek (ahol erre a feladat módot ad) • csak egy-két kisebb hiba fordul el•, melyek a megértést nem befolyásolják	Emelt szint• feladatokban • pontos és változatos szóhasználat	Emelt szint• feladatokban • jól felépített szöveg • a szövegkohéziót biztosító eszközök használata változatos • a szöveg bekezdésekre tagolása és elrendezése teljesen megfelel•
6				
5	• a vizsgázó a feladatot megoldotta, minden tartalmi egységr•l írt • a szöveg stílusa és a regiszter alkalmazása nem teljesen következetes a szöveg jó benyomást tenne a megcélzott olvasó(k)ra	• helyes nyelvtani szerkezetek és jó helyesírás, kisebb pontatlanságokkal • változatos nyelvtani szerkezetek (ahol erre a feladat módot ad) • A bonyolultabb szerkezetekben el•fordulnak hibák, de ezek nem befolyásolják a megértést	• változatos szóhasználat	• világosan felépített szöveg • a szövegkohéziót biztosító eszközök használata megfelel• • a szöveg bekezdésekre tagolása és elrendezése többnyire megfelel•
4				
3	• a vizsgázó a feladatot elkezdte, de nem oldotta meg teljes egészében, az adott feladathoz az Értékelési útmutatóban meghatározott számú tartalmi egységr•l írt • a szöveg megfelel• benyomást tenne a megcélzott olvasó(k)ra	• megfelel• nyelvhelyesség és helyesírás • a nyelvtani szerkezetek még elfogadhatóan széles skálája • több hiba fordul el•, de ezek nem akadályozzák a megértést	• odaill• szóhasználat	• megfelel•en felépített szöveg • odaill•, de gyakran csak egyszer• szövegkohéziót biztosító eszközök • a szöveg bekezdésekre tagolása és elrendezése általában megfelel•, de következetlenségek el•fordulnak
2				
1	• A vizsgázó nem megfelel•en fogott hozzá a feladat megoldásához, a legtöbb tartalmi egységr•l nem írt • a szöveg rossz benyomást tenne a megcélzott olvasó(k)ra	• nem megfelel• nyelvhelyesség és helyesírás • a nyelvtani szerkezetek sz•k skálája • a gyakori hibák gátolják a megértést	• sz•k skálán mozgó és/vagy oda nem ill• szóhasználat	• nem megfelel•en felépített szöveg • nagyon kevés vagy többnyire nem megfelel• szövegkohéziót biztosító eszközök • a szöveg bekezdésekre tagolása és elrendezése nem megfelel•,
0	a vizsgázó el sem kezdte a feladatot, VAGY hozzákezdett, de a szöveg tartalma nem megfelel•, VAGY olvashatatlan a kézírás	még a legegyszer•bb szerkezetek is teljesen hibásak, VAGY túl kevés a szöveg az értékeléshez	egyáltalán oda nem ill• szóhasználat, VAGY túl kevés a szöveg az értékeléshez	a szöveg egyáltalán nincs felépítve, és nincsenek szövegkohéziót biztosító eszközök, VAGY túl kevés a szöveg az értékeléshez