Chapter 3

EXPERIENCING THE EXAMINATION DESIGN, CONTENT, MATERIALS AND PROCEDURES[1]

Judit Kiss-Gulyás

This chapter examines how course participants experienced aspects of the examination. It presents the procedures used, how teachers view the current examination and how the new examination specifications were received. It also shows how the participants responded to new examination task and text types, and to results from the pilot examinations.

3.1 Introduction

I was involved as trainer in the second pilot course (held in Debrecen), conducting 50% of the sessions[2]. The course, which was in several ways different from the first pilot course in Eger, aimed to inform teachers of English from Hajdú-Bihar County (mostly from Debrecen) about the current state of affairs concerning the proposed school-leaving exam, and to prepare teachers' minds for the radical changes the exam would have on the teaching and learning of English in secondary schools. This aim was meant to be achieved in such a way that participants should have a positive view of the proposed examination.

The materials and procedures for the pilot courses had been prepared over the preceding 18 months by the INSET team of the Examination Reform Project, and had been first piloted in Eger. After the Eger course a general revision of the materials and procedures was carried out, and a Course Booklet was produced for the second pilot course. Unlike the Eger course, by the time of the Debrecen course the piloted exam tasks and the results of the pilot studies were available (see Alderson *et al* 2000: 258-278). This fact added further potential to make the most out of the course: as the piloted tasks were available for further trailing, participating teachers were invited to try out the tasks with their own students. This was part of the home assignments (see Chapter 6 for details), and enabled teachers to investigate the tests in action. Also, they were asked to take notes of their learners'

[1] Personal note: The chapter title refers to the experience course participants had with the exam content and materials. For me it also refers to my own experience: I was not involved in the design of the course, neither did I know too much about the proposed examination when I was asked to participate as trainer in the in-service course. So I investigated the course materials and also the proposed examination both with curiosity and interest, and also with critical eyes. I made several suggestions, most of which were appreciated by the course designers. Perhaps the' innocent eye' observed issues that remain unnoticed by those who had been working on the documents for a long time.

[2] I ran the sessions on reading, writing and use of English, so my examples will be take from these areas.

reactions and their own impressions. The results were recorded and compared to the achievement of those groups who were involved in the piloting part of the programme. It was a very important component of the course, as we will see. Teachers really appreciated the fact that they had the opportunity to see what the proposed exam tasks might be like, and to critique the tests.

3.2 Procedure used in the course to enable teachers to experience the likely new examination

Teachers were exposed to the proposed examination in a well-thought out, careful way (see the underlying principles of the 'learning process' in Chapters 2 and 3): the tasks in the sessions and also the home assignments focussed on aspects of the proposed exam, approaching it holistically first, and then teasing it apart.

The concrete steps were the following:

First, participants were asked to express their opinion about the current school-leaving examination. Second, they were asked to study the Working Document (Vándor *ed*, 1998), which contains the description and specification of the new examination. This step enabled participants to familiarise themselves with, and critique the already formulated draft documents about the nature of the exam, its components and the possible text and task types. The third step was that participants were asked to look at the course books they were using in their respective schools, and the classroom procedures they were employing. The aim of this step was to make teachers investigate to what extent their classroom practices matched the exam task and text types, to see if there was a gap between current classroom practices and the exam content. As a fourth step, teachers were asked to try out some of the sample, already piloted, tasks with their own learners. Results were discussed and evaluation followed. The final step was to look at the classroom implications: suggestions were made as to what kind of classroom activities would help students prepare for a given component of the proposed examination.

These activities all contributed to raising teachers' awareness about the intricate relationship between testing and teaching, and also to developing in them a positive attitude towards the examination.

Let us now look at these steps in more detail.

3.3 Current exam, its currency and validity – opinions about the current school-leaving examination

One of the initial aims of the second pilot course was to find out what practising teachers think of the present school-leaving examination (*nyelvi érettségi*). It was envisaged by course designers that there would be more problems than positive opinions about the current exam, thus the need for a change would be a logical conclusion.

Participants were given ample time to collect thoughts and formulate opinions as there was a pre-course task which aimed to raise teachers' awareness of the features of the current examination. Furthermore, participants were asked to think about their

current classroom practices, and the relationship of classroom practice to the school leaving examination (pre-course tasks 1 and 2). In the first session the opinions were shared and there was a general consensus that changes were necessary. This conclusion, made by course participants, corresponds with the message both of other practising teachers and testing experts: studies (e.g. Fekete *et al*, 1999) have appeared in the past couple of years which critiqued the current language examination, arguing that it has to be replaced by one which responds to practical needs and expectations, and is theoretically well-founded. The idea of replacing the current exam was accepted unanimously by course participants.

Table 1 summarises the comments participants first made about the positive and negative features of the current school-leaving examination. Opinions are grouped from the point of view of the teacher and learner, and then the test features mentioned by course participants are enumerated.

Table 1: Participants' comments on the current school-leaving examination

	Positive features of the current school-leaving examination	**Negative features of the current school-leaving examination**
From the point of view of practising teachers	• Familiarity with procedure (basic task types, levels, grading) • Availability of preparation materials • High percentage of passes (the number of classes and the test requirements and tasks match) • Predictability of the oral part of the examination (topics are given, students can systematically prepare for them) • Predictability of results (the assessor is the teacher of the given pupil)	• Unpredictability of task and content of the written test (in terms of vocabulary and structures) from one year to the next • Level of the test is unequal • Assessment of the oral examination is subjective, no guidelines for marking • Assessment of the written part is problematic (no detailed key, assessment scale, and some test items may generate more than one good solution) • Predictability of the content of the oral examination: no real communication
Learners' views	• Familiarity of topics for the oral examination • Success in most cases. Pass level is very easy to achieve. • Classroom teacher is the assessor (mutual assistance)	• Useless in terms of practical use (not appreciated by employers, universities) • Due to unpredictability of task types in the written examination it is difficult to prepare for it. • The translation part is difficult.

Test	• Familiar • Relatively cheap to prepare • Easy to administer • Procedures are clearly worked out and known nation-wide	• No currency (acceptance, or market value is low) • No specifications • No clear identification of levels • No clear guidelines for grading • Lack of validity (content) and reliability (source: scoring) • Not compatible with the internationally standardised tests • Tasks are not piloted • Lack of objectivity • Separation of testing and teaching • Not all skills are tested

3.4 What course participants knew

Most language teachers knew that the current school leaving examination had no currency: it was not regarded as a suitable and accepted entry pass into the job market or higher education. Its motivational power was minimal and it did not satisfy some of the basic principles of language testing: it was not reliable, it lacked validity and it had a negative washback effect on classroom teaching. To be fair, one has to acknowledge that some changes had been made in the past few years, shifting the examination towards a somewhat more communicative exam (see Ábrahám and Jilly in Fekete *et al*, 1999: 21-53). These slight improvements, however, were unaccountably introduced, and there was no guarantee that the text and task types of the previous year's test would re-emerge in the following year's test. Thus preparation for it, in terms of sensitising students to some task types, was difficult. There was always a surprise element in the test, making students and teachers dissatisfied and disillusioned.

One may argue that the need for a change was as much acknowledged and even initiated from bottom-up as from 'above' by higher level decision makers: practising teachers seemed to accept that the current '*nyelvi érettségi*' had to be reformed. How they envisaged the reformed exam varied but very often the model was either the 'currency-holder' Rigó utca examination, or some standardised foreign language examinations (e.g. Cambridge, Oxford exams).

3.5 Description of the proposed exam – Opinions about the Working Document available at the time of the INSETT course

Course participants received the most recent drafts concerning the likely new examination in the form of the *Working Document for the School-leaving Examination in English, Intermediate and Advanced Levels* (Vándor *ed*: 1998). The aim was to enable them to thoroughly investigate the document and to have them experience the concept and content of the proposed examination. Pre-course task 3

(Course Materials Package, p. 8) encouraged teachers to formulate questions, express queries, fears, or hopes in connection with it.

Let us see what course participants thought of the *Working Document*. First of all, they really appreciated that their opinions were considered. They took the evaluation of the document seriously and came forward with relevant and interesting comments. The issues participants raised show striking similarity with the points and observations made by teachers who participated in the piloting of the exam tasks (for details see Alderson *et al*, 2000).

The natural way of approaching the proposed exam was to compare it with the 'old', current examination, and this is what course participants did consciously or subconsciously. Table 2 summarises participants' first reactions.

3.5.1 What teachers experienced first

Table 2 Comparison of the current examination with the proposed examination

	Current examination	**Proposed examination**
Test construction	No central control	Careful specifications
Task types	Unpredictable, varied	Predictable, level-dependent
Text types	Variation	Authentic
Piloting the test items	No piloting	Items have to be piloted
Levels	8 different versions, levels	2 levels (intermediate, advanced)
Grammar	Tested, often at sentence level	Not tested separately, Use of English (text-based resting of grammatical knowledge) at advanced level only
Translation	Exists	None
Skills • **Reading** • **Listening** • **Writing** • **Speaking**	 In the oral examination, and in the written examination. The reading component is not meant to measure the subskills involved in reading None Mostly filling in words, paraphrasing, no guided writing tasks proper No authentic communication, predictability of topics	 At both levels, aim: to check Ss' reading subskills in a more precise way Exists Guided writing tasks appropriate for the level Authentic interaction, level dependent topics and task types
Assessment and evaluation • **Written part** • **Speaking**	 centrally provided key to multiple choice tests, no key to translation impressionistic grading	 centrally provided key +detailed analytic rating scale to grade Ss' writing analytic rating scale

3.5.2 Feedback from participants– fears, doubts and hopes

It became clear from the Working Document that the likely new examination was radically different from the current school leaving examination: it is a two-level (intermediate and advanced) skills-based language examination with several novel elements: translation is out; the materials used for testing reading and listening are all authentic; the tasks are real and life-like. Grammar (in its traditional sense as a separate paper) is out, at least at intermediate level, and listening is incorporated into both levels. The draft examination documents clearly specify the text and task types for the two proficiency levels. By the very nature of this classification, there are text types that can be integrated into an advanced and intermediate test with equal ease. As the task types were identified uniformly for the two levels participants felt that there was not much difference between the text types an intermediate or advanced student is supposed to be familiar with. Furthermore, the treatment of speaking appears to be radically different from the current examination: the speaking component is envisaged as a paired examination with external assessors. Also, an analytic rating scale is introduced in the assessment. An analytic rating scale is also used for the evaluation of the writing tasks. These novel elements generated numerous questions.

Practising teachers' initial opinions and feedback on the content and materials of the examination reflected their well-founded hopes and justifiable fears. Participants' fears were related to learner achievements mostly: they were afraid that more students would fail than at present; they were worried about the achievement of lower ability students. The second major point participants raised concerned the implementation and administration of the test components, especially the listening and speaking parts: it was not clear to participants for example, how the spoken examination would be organised, and who the assessors might be. Also, they wanted to hear more about how the testing of listening would take place. A further issue raised was the weighting and measurements of the subcomponents. Almost all of them worried that the number of English classes per week (in most schools) might not be enough to prepare students for an examination of the proposed type. There were questions relating to the definition and identification of levels, more precisely: who would count as an intermediate and an advanced student. Their hopes were related to the prestige and acceptability, or currency of the new examination, and its positive effect on classroom practices.
This general discussion was followed by the detailed investigation of the different components of the proposed examination. Let us see what participants thought of the text and task types for testing the skills. Of the five components three will be dealt with: two skills-based components (one of which tests comprehension, reading and the other production, writing), and a knowledge-cum-skills-based component[3], use of English.

[3] This term attempts to mirror the treatment of 'grammar' in the proposed examination: grammar in this test does not mean sentence-level, prescriptive knowledge of the rules of the language. Rather, the Use of English component aims to test candidate's knowledge of not only the formal properties of grammar but also the function-and meaning components of a given grammatical phenomenon, beyond the level of the sentence. As the tasks are text-based, some of the reading subskills are also involved. Thus the term. Note that this component appears in the Advanced examination only.

3.5.3 Feedback from participants before trying out the proposed examination tasks and after piloting them

In this part we will see what teachers said about the relevant parts of the Working Document and then about the tasks they were asked to try out with their own groups. Note that as the course was designed around the concept of discovery learning, participants themselves were invited to solve some of the tests. The procedure was similar in the case of the different components: first the *Working Document* was investigated, then the proposed tasks were compared to the tasks in widely used modern course books. Thirdly, participants were asked to critique the exam tasks before and after trying them out. Let us see what came out of the discussions.

3.5.4 Reading

Before analysing the relevant part of the *Working Document* participants were invited to discuss the order and importance of the four skills in their own practices. What determined the place of the individual skills in their language programme depended, as it turned out, on how the respective course book treated them, and also, how the teacher looked at them. Teachers seemed to agree that reading as a skill is usually appropriately dealt with in the course books they are using. After investigating current classroom practices and also the generally accepted theoretical claims as to the subskills involved in reading, the tasks and text types listed in the examination document were looked at and commented on.

Initial comments made by participants centred around the novel elements: for example, texts are all taken from authentic sources; the text types try to mirror the text types we encounter on a daily basis when we read in the L1 or in the L2; the number of test items and their ratio in the total score in percentages are given; the length of text is given; no dictionary use is allowed.

The reading component was designed for two levels, intermediate and advanced. The task types appeared to be identical for both levels but the texts were level-dependent. Although texts were meant to be level dependent participants had the feeling that even the intermediate level students were assumed to be able to cope with a wide variety of text, including reference books, encyclopaedias, literary texts, DIY books. The teachers' view was that it might be too demanding as some of the texts use special vocabulary. It must be added here that the document itself states that candidates are not expected to understand every word in the text. The presence of unfamiliar words requires different text-attack skills from the learners: they are encouraged to tolerate ambiguity and to deduce meaning out of context. The immediate classroom implication teachers formulated was that more attention should be given to the teaching of different text-attack strategies.

As for the task types, course participants indicated that most of them were familiar, and also that the course books they were using contained tasks of the same type. There were few task types which participants themselves did not have a precise understanding of (for example, multiple matching and banked gap filling), and these were elaborated on. As for test administration, some of the participants were against the idea of giving the instructions in English exclusively.

The general discussion of the appropriate part of the *Working Document* led to hands-on experience when participants were asked to look at three of the piloted examination tasks and analyse them with the help of the following set of tasks (Course Materials Package p.8).

a. Look at the examination document and identify the text and task types.
b. Decide which of the reading skills we looked at in Session 1 are needed to complete the task.
c. Give your opinion of these tasks. How difficult are they? Are they suitable for any of your students?
d. Are they interesting for students?
e. Report your views to the whole group.

Identification of text and task types on the basis of the *Working Document* proved to be less demanding than the identification of the subskills involved in the individual tasks. Group discussion was extremely useful as the more experienced and theoretically better-trained participants helped the others.

3.6 Comments on sample texts and types[4]

3.6.1 Reading task one – Fatherhood transformed me

This text consisting of 370 words was a magazine article about the life of a TV personality. The subskill it aimed to test was understanding relations between parts (text cohesion and coherence). The task type was multiple matching: candidates had to find which paragraph should be put into the gaps in the text. The gaps were created by removing 5 paragraphs from the text. These paragraphs were presented to the test taker in a jumbled order. There were five items in this test.

The general opinion was that students might have problems with this task. In actual fact teachers themselves took a good amount of time over it. It may have been attributable to technicalities (e.g. layout design: the parts of the test were on two pages), to unfamiliarity of task type, or to lack of enough practice.

3.6.2 Reading task two – Advertisements

This text of 200 words contained 9 small ads or classified ads from newspapers and magazines. The subskill it aimed to test was understanding gist, and separate the relevant from irrelevant information. The task type was multiple matching: candidates had to find which heading fitted the short texts. The gaps in this task were created by taking out certain pieces of information: these removed parts referred either to the name of the advertising company or to the advertised product. Authenticity was retained in that the layout was kept. There were eight items in this test.

The general opinion was that the task was easy, as the removed words and/or company names could be found in exactly the same form in some of the ads. This initial thought of participants about the difficulty level of the second reading task

[4] For more details about the reading tests used for piloting see Cseresznyés in Alderson *et al* (2000: 160-178). Bear in mind that the second pilot course was held in 1999, so the information presented in Chapter 13 by Cseresznyés was not available.

was further reinforced by the results of the pilot test, in which the mean score was 80%. The task, which was originally designed for basic level[5], was judged as easy by most participants.

3.6.3 Reading task three – Tadpoles

This text of 32 words was an encyclopaedia entry, describing a process. The subskill it aimed to test was to find specific information, and again, to separate the relevant from non-relevant information. The task type was multiple matching, pictures had to be matched to texts. The number of test items was 5.

Participants argued that though the text was very short it contained words that elementary level students, the target population of this test, are not really familiar with. Pictures and knowledge of the world may help, but then one can question whether it is a good test of reading or not.

3.6.4 General comments made by participants on the three pilot tasks

The following initial general comments were made on the three tasks:

- Multiple matching was used in all three tasks despite the fact that this task type was not widely used in the classroom, and thus was not really familiar to teachers, and most probably to learners either.
- The layout of the texts (size of rubrics, clarity of instructions) was found unsatisfactory. There was general agreement the layout had to be improved.
- Some participants argued that the instructions should be given in Hungarian.
- It was felt that the difficulty level of test items within tests varied considerably.
- Some items were too easy to discriminate (e.g. Item 5 in Task 3, which was done by 94% of the test takers).

3.7 Opinions about the reading tasks after trying them out

The reading session follow-up task was that participants were asked to try out the tests with the groups, or one of the groups of students they were teaching.
The following procedure was suggested:

a. Record students' marks in the given table.
b. Ask students for some verbal and written comments on the task.
c. Summarise their responses and write the summary in the table.
d. Compare your students' comments with the comments you made about the tasks.

This had to be compiled for the last session on reading, for Day 3 of the course. Note there was almost a month to do this task. Note also that participants were free to choose one task out of the three, and there were some participants who tried out all the three tasks.

[5] Note that the first version of the proposed examination document was designed for three levels: in addition to the intermediate and advanced levels there would have been basic level. For this reason some of the piloted tasks were basic level tasks, for example, Reading task 3 (Tadpoles).

The last session on the reading component of the proposed examination centered around giving feedback on the tasks that participants themselves had tried out with their own students. The procedure was the following: first student performance was looked at and participants compared the results of their pupils with the results of those who participated in the trialing. Second, students' comments were investigated on the individual tasks.

Let us now look at the tasks again, following the procedure described above.

3.7.1 Reading task one – Fatherhood transformed me

3.7.1.1 Students' performance

This task was tried out in 14 groups (number of students: 181) both in grammar school and vocational school environments. The total score (in percentages) shows considerable variation between 25% and 88% with an average of 52%. Table 3 shows the mean score of the 181 strong population on the individual items.

Table 3 Reading task one – Mean scores on the individual items in %

Item	Mean scores (in %)
1	61
2	54.7
3	53
4	47
5	45

3.7.1.2 Comparison of classroom results with the results of the piloting

As for the difficulty level of Task 1, there was general agreement that it was the most difficult of the three. In this respect the findings correlate with the results of the pilot study, which also found this text and task type the most difficult. The mean score (52%) is higher than that of the pilot (29%) (cf. Cseresznyés in Alderson *et al*, 2000)[6]. Item 5 proved to be the most difficult with Item 4 as a close follower, and Item 1 the easiest. In the pilot Item 3 appeared to be the easiest and Item 4 to be the most difficult.

Note that one has to look at these results carefully for different reasons: first, multiple matching tasks are difficult to measure as the initial choice, which may be wrong, may have an effect on the rest of the test. Second, there were considerable variations within and across groups, which would require detailed analysis. That is, however, not within the scope of this chapter. Intuitively, the variation can be attributed to the heterogeneous nature of the test taking population both in terms of level, age and exposure to language.

3.7.1.3 Comments on the task and text

Most students and teachers agreed that the task itself was interesting and enjoyable. There were negative opinions as well: some students found the text boring and

[6] Note that the pilot test scores were not given to course participants before they administered the tests as part of their home assignment)

uninteresting, others mentioned that the task type was unfamiliar. As for the vocabulary level, it was found appropriate. There were comments concerning the instructions: students found the wording complicated and the layout (boxes and numbering) confusing. The time students allocated to the task varied but for most it took around ten minutes.

3.7.2 Reading task two – Advertisements

3.7.2.1 Students' performance

Ten groups (number of students: 133) tried out this task both in grammar school and vocational school environments. The total score (in percentages) shows less but still considerable variation between 57% and 97.5% with an average of 84.5%. Table 4 shows the mean score on the individual items.

Table 4 Reading task two – Mean scores on the individual items in %

Item	Mean scores (in %)
1	86.3
2	83.9
3	81.2
4	73.7
5	89
6	94.3
7	91.2
8	72.7
9	88.5

3.7.2.2 Comparison of classroom results with the results of the pilot (Pilot: 80%)

In the case of task two the two means are similar: 84.5% was the mean and the pilot mean was 80%. The high score indicates that this text and task type is easy for the students involved.

3.7.2.3 Comments on the task and text

Most students found the task easy and fast to do. Few students had task-related problems.

There were problems with the instructions in that students found them complicated, the numbering and the boxes confused them. The task took about 10 minutes or less to do.

3.7.3 Reading task three – Tadpoles

3.7.3.1 Students' performance

This task was tried out in 4 groups (number of students: 54) both in grammar school and vocational school environments. The total score (in percentages) shows little variation between 68% and 88% with an average of 77.8%. Table 5 shows the mean score on the individual items.

Table 5 Reading task three – Mean scores on the individual items in %

Item	Mean scores (in %)
1	81.6
2	76
3	64.6
4	70.2
5	100

3.7.3.2 Comparison of classroom results with the results of the pilot

Here again the pilot results are much the same: classroom mean is 77.8% and the pilot mean is 76%. As is apparent, this is a relatively easy task for the given population. The easiest item is Item 5, which produced 100% correct answers. In the pilot Item 5 was also very high, which means that the item does not really discriminate well, it is too easy.

3.7.3.3 Comments on the task and text

Students found the task straightforward and easy, but they also added that the text relied heavily on one's knowledge of the world. One student wrote that 'those who are not interested in Biology might have a hard time with it.' Others argued that the vocabulary level and test level did not match. Most found the instructions clear and unambiguous. Students took less than ten minutes to do the task.

3.7.4 Summary

As we have seen, the results course participants reported showed similar tendencies to the pilot study results: there seemed to be agreement about the general difficulty level of the individual tasks and texts, and the difficulty level of the individual items also showed similarities.

Recall that initially almost all course participants critiqued the layout and the organisation of the test questions and numerous comments and suggestions were made. On the basis of post-trial reports it can be stated that some of the initial fears of course participants appeared to be unfounded.

3.8 Writing

It came out very strongly in the introductory activity on the four skills that writing is a skill that is neglected in the classroom, and that most teachers, under the pressure of circumstances cannot devote enough time to it. Classroom writing is often consolidation of the oral work that has been done in the given lesson, and often takes the form of copying. Composition writing is a take away activity, and teachers are only interested in the final product not the process of writing. This situation is very similar to what Nikolov in Fekete *et al* (1999) reported.

As writing seemed to be a problematic skill, more time was devoted to it in the course than to the other skills. The increased number of sessions enabled participants to think critically of what it means to say that writing is a

communicative act, and via the discussions the importance of the unity of topic, purpose and audience came out very clearly.

In the next writing sessions the task and text types listed in the *Working Document* were investigated. The following general comments were made:

- Candidates have to write letters in response to some input, which is new.
- The text types the input texts represent are realistic.
- The task types show variation.
- Candidates are supposed to write a text of a transactional nature and a text, which allows free expression.
- Degree of guidance depends on level: intermediate students get more guidance in terms of content, organisation and language.
- Unfortunately the *Working Document* does not specify clearly which task types are more likely to occur at the given levels. (Vándor *ed*, 1998: 18).
- Dictionary use is not allowed.
- Candidates' performance is assessed by using an analytic rating scale.

The next task for participants was to look at three of the already piloted writing tasks[7] and identify the text and task types. Also, participants were asked to comment on how well their students were likely to perform on these tasks.

3.8.1 Comments on sample texts and types

The following three piloted writing tasks were shown to participants:

Writing task 1– students were required to create a text (a formal letter of 70 words) on the basis of a verbal prompt (newspaper advertisement).

Writing task 2– here again students were required to create a text (an informal letter of 80-100 words) on the basis of a verbal prompt (diary entries).

Writing task 3– students were required to write a text (a formal or informal letter of 100 words) on the basis of a given situation (without additional prompts).

3.8.2 General comments made by participants on the three pilot tasks

Participants found the following aspects important to mention:

- The tasks are realistic in terms of topic and purpose and they are 'communicative'. It is clear who the audience (reader) is, what the relationship is between reader and writer (which determines the level of formality one has to use).
- Appropriate guidance is given to students, the less creative ones can also demonstrate their ability to write in English.
- Limiting text length in words may cause a problem, students are not used to it.
- In the case of Tasks 1 and 2 standardisation as regards the answer is possible, but in the case of Task 3 it is not possible.

[7] For details of the piloted tasks see Szabó *et al* in Alderson *et al* (2000: 100-122).

- As dictionary use is not allowed, more emphasis should be placed on teaching vocabulary in the classes.
- The introduction of the analytic scale will add objectivity, as so far there had been no guidelines for markers.

3.8.3 Using analytic rating scales for evaluating student writings

The other aim of the sessions on writing was to call participants' attention to grading reliably. As teachers believe that they usually grade reliably (with the system they may have worked out for themselves), it had to be demonstrated with the help of an awareness raising activity that it was not necessarily the case. Participants were asked to grade the same compositions impressionistically twice, with some time lapse in between without knowing that they would have to grade the piece the second time. It was well demonstrated that the same teachers' judgement on the same composition differed on different occasions, and also that there was a difference between how they and others evaluated the same piece of writing. This activity was well received and convincing: teachers' awareness of subjectivity involved in marking the written components increased, and they realised the necessity of using a detailed analytic writing scale, which definitely increases scorer reliability.

The awareness raising activities clearly showed that subjectivity of marking was often observed if teachers used impressionistic grading. Also, not only intra-rater reliability, or individual inconsistency was raised but also inter-rater agreement. With the help of hands-on tasks it was shown that participants approached the same piece of writing differently, which simply means that there was place for subjective judgements. All these observations led to the conclusion that in order to increase marker reliability and objectivity analytic rating scales should be introduced. These would enable practising teachers to evaluate pieces of writing more objectively. Participants were exposed to an analytic writing scale, and were asked to comment on it. Table 6 shows teachers' opinions.

Table 6 Participants' opinions about using an analytic rating scale

Positive features	**Queries, problems**
Detailed guidelines for grading are necessary. It will definitely improve scorer reliability and minimise subjectivity. There is a clear shift of focus: the rating scale focuses on the communicative effectiveness of the piece of writing, not only on its formal aspects. Double marking of papers is essential. Marker training is important.	The use of analytic scale will increase grades, which is very promising for less bright students. The guidelines are too lengthy and vague at places (e.g. What is meant by 'viszonylag', 'jobbára'; What is meant by serious mistakes at intermediate and advanced levels?). Sample solutions should be provided. Points should be subtracted for very short pieces.

3.8.4 Opinions about the writing tasks and using analytic rating scale to evaluate student writing after trying them out

The writing session follow up task was again to try out one of the sample writing tasks with one or more groups of students and apply the rating scale. This was done in such a way that students themselves were exposed to the scale before they were asked to do the task. Then teacher and learner comments were collected and compared.

Though teachers were afraid of using the analytic scale, they all agreed that it was not as complicated as they had thought it would be, and after some practice one could easily get used to it. All acknowledged that the extra time and energy were worth the effort as the evaluations were more objective. Also, students liked the idea of the teacher using an objective scale for grading. When being exposed to the scales students were confused by the set of criteria teachers might use for the evaluation of their pieces of writing.

There were several critical remarks made by participants as well. These did not question the validity and importance of an analytic rating scale but emphasised that clear wording and categories were crucial. Also, fuzzy labels that lend themselves to several interpretations have to be avoided[8]. In the scale the following categories were used: communicative effectiveness, or task achievement (*kommunikatív cél teljesítése*), vocabulary *(szókincs gazdagsága*), accuracy, spelling (*nyelvhelyesség, helyesírás*) and organisation *(a szöveg megszerkesztettsége*). There was general agreement that the task achievement and organisation components were easier to evaluate than the other ones. Assessing vocabulary and grammar appeared to be more problematic, due partly to the above-mentioned vague terminology (e.g. interpretation of 'fairly large range of vocabulary', or 'limited range of vocabulary'). Some teachers liked the idea of writing the analytic scale in Hungarian.

Let us see what some participants wrote about the scale:

'It is very good to have rating scales like this because teachers usually grade their students according to grammar and spelling and most of the time students write easier and simpler sentences to avoid mistakes. It was difficult to differentiate between the stages, whether a student should get 5 or 6 points for a given component.'

'I was a bit scared at the beginning as I had never used such a scale before but after reading it thoroughly I found it very interesting. At the beginning it seemed complicated – and of course I do not intend to say that rating is simpler now – but it seems to me that I can be more objective. Sometimes it is difficult to decide between the two grades but it may be closer to reality. If we regularly grade the students' writing according to this rating scale, we will get used to it and it is much better to have a detailed guideline than to have almost nothing at all.'

'I do not think we can make a really objective rating scale. I did the task twice at two different times and I had different results. It seems to be useful but the harder you try the more complicated it gets. I may need some further practice to work out certain standards.'

'I had to read each criterion several times and it was rather difficult to decide whether the given piece of writing fits into the different categories at all, or which definition it

[8] Note that the version of the analytic rating scale participants worked with has since been replaced by a new, more detailed one, taking into considertation the suggestions made by participants (see Szabó *et al* in Alderson *et al*, 2000: 100-122).

could be matched with. I think the criteria should be formulated more carefully and accurately.'

On the whole, it was found that teachers who were sensitised to the use of the scale in the sessions generally found it comfortable and safe to use in the assessment of their learners' written pieces.

3.8.5 Comments on the writing tasks

The students liked the tasks, they liked the prompts, which offer even the less imaginative candidates some guidance and help. Some students did not like the idea of writing compositions without dictionary use. Others found the word limit too restrictive arguing that the compositions were too short for them to demonstrate their knowledge of the language. They also expressed their fear of having an external examiner.

Teachers thought that their students needed further input on the formal side of letter writing, and also on the use of cohesive devices.

Originally, student achievements on some of the tasks would have been compared to the pilot results. This will not be part of the present account for the following reasons: first of all, different teachers tried out different tasks and from their homework assignments it is not clear which. Second, as a different analytic scale was applied from the one used in the evaluation of pilot tests, comparison appears to be difficult.

3.8.6 Summary

Students were able to do the three tasks at their competence level, and they found them interesting and realistic. Course participants also found the tasks fair, and the analytic scale very useful.

3.9 Use of English

A component on testing students' grammatical knowledge has always been part of the school leaving examination. One may argue that the Use of English part of the proposed examination is the 'leftover' of the traditional grammar component but a completely different approach to it is applied here.

The detailed requirements state that 'the purpose of the Use of English paper is to assess whether the candidate possesses the lexical, grammatical, semantic and pragmatic knowledge that will enable him/her to communicate independently. Contrary to the specifications for school-leaving examinations in other foreign languages this paper constitutes a part of the English exam at advanced level only' (cf. Detailed Requirements) Linguistic competence will be tested both at sentence and text levels. Sentence-based tasks will include multiple choice items and sentence transformations. Text-based tasks will be the following: gap filling, banked gap filling, multiple choice, error identification, identification and correction of errors, arranging jumbled texts. Grammar is looked at as a support system, from which it follows that Use of English is approached from a discourse perspective:

items may be sentence- or text-based and also the task types will be both sentence-based and text–based. It is emphasised that even in the case of sentence-based items the source of the item should be an authentic text.

Such an approach to testing Use of English has the consequence that there may be some degree of overlap between reading and Use of English part. More specifically, some of the subskills that are meant to test reading also test Use of English. The main difference as Martsa and Nyir• (in Alderson *et al*, 2000) argue is that tasks of the reading component test recognition of grammatical features, whereas the Use of English component requires the production as well as the recognition of the structures.

Relatively few, three sessions out of 50 were devoted to the investigation of the Use of English component of the proposed examination. In the first session teachers were invited to express their views about teaching grammar. There was general agreement that teaching grammar in the communicative way, which means that not only the form but also the meaning and function of a given structure are taught, is difficult but necessary. Participants also argued that it can be time-consuming to prepare students for grammar tests despite the fact that good materials are available.

3.9.1 General comments made by participants on the three pilot tasks

The *Working Document* was looked at in terms of text and task types. Then three piloted examination tasks[9] were shown to participants, who had to identify the text and task types the sample tests represented, and also they had to define which language area/s were tested in the given task.

The three tasks were the following:

Task A (text-based) – What on earth

The authentic text consisted of 209 words and 15 items were designed for it. Visual prompts were also provided. The task type was gap-filling. The investigated language area was the use of prepositions.

Task B (text-based) – Spice girls

The authentic text consisted of 225 words and 16 items were designed to go with it. The task type was gap-filling. The language area tested was vocabulary, with particular emphasis on the cohesive devices.

Task C (sentence-based) – Number of items: 10

This task consisted of 10 'isolated' sentences with some parts underlined. The task was to identify the underlined part, which is incorrect, and correct it. Note that of the underlined parts only one is incorrect. Of the 10 items only three focussed on structure (7, 9, 10), and the rest was mostly lexis or a mixture of lexis and structure.

[9] See Martsa and Nyir• in Alderson *et al* (2000: 179-197) for more details.

3.9.2 Comments on the task and text after trying them out

3.9.2.1 Task A – What on earth

This text-based task was done by eight groups (number of students: 106). The average score (in %) was 28%, ranging from 0 to 60%. The overall impression of teachers was that they themselves found this task difficult. The fact that pictures were provided to help testees figure out what was missing from the text did not seem to help at all. The vocabulary was 'unfair', as one participant remarked: there were a lot of unfamiliar words in the text. Bearing in mind that the task intended to test students' knowledge of prepositions and cohesive devices it was a huge disadvantage. Things were further complicated by the fact that there were items where more than one good solution was feasible (e.g. 3, 9 and 14). Most teachers did not try out the text, as they did not have a group that could be labeled as advanced.

Comparison of classroom results with the results of the pilot

The mean of the pilot on this task was very similar, 32% to the mean of the classroom learners (28%). Also, the same items appeared to be problematic and were reported to be difficult: Items 1, 3, 7, 9 and 10.

3.9.2.2 Task B – Spice girls

This task was done by six groups (number of students: 76). The average score (in %) was 27%, ranging from 0 to 68%. There were several initial comments participants made. Although the task and text were found interesting, some critical remarks were made, which concerned the vocabulary load and the cultural implications of the text. Teachers found the text difficult. Most students found the task demanding for the following reasons:

- The vocabulary was too difficult.
- The text was culture-bound, some background knowledge was definitely needed.
- Distribution of gaps appeared to be problematic: there were so many items that tested cohesive devices and were deleted that the text lost its integrity, which made it impossible to reconstruct it at places.

Comparison of classroom results with the results of the pilot

The pilot tests showed that this test was beyond the competence level of the target student population (mean: 14%). Participants in the second pilot course produced better results but the mean score is still low (27%).

3.9.2.3 Task C (sentence-based)

This task was done by eight groups (number of students: 100). The average score (in %) was 28%, ranging from 0 to 60%.

Most students found the task very difficult, some wrote that it was confusing and frustrating. There was a general agreement that they would not like to have such a task in the final examination.

The difficulty of the task derived from different sources: First, students were not really familiar with this task type, it is not done frequently enough. Second, although this task is meant to represent an advanced level task it was done by students who represent the intermediate, or even lower levels, due to lack of

availability of advanced level learners. A smaller problem, which was also mentioned, concerned the layout of the task. For the successful solution, as one test taker argued, an excellent command of English is required. Others wrote that the text used very difficult vocabulary items. Interestingly enough, students were better on the items that tested structure exclusively than on the ones testing lexis, or both lexis and structure. As the text contained vocabulary items students did not know, it was difficult to identify lexis-related errors.

Comparison of classroom results with the results of the pilot
Pilot test results (mean: 27%), which are almost the same as the classroom results, also showed that this task seemed to be too difficult for this population. Investigation of the individual test items also reveals similarity between the two groups: both student groups found Items 7 and 9 easiest in the text, the items that test structure.

3.9.3 Summary

Participants found the Use of English component difficult and some teachers went as far as to say that they themselves felt unsafe in some of these areas. They demanded training on Use of English so that they can prepare their students for this component.

3.10 Conclusion

It has been shown in this chapter that course participants were offered ample opportunities and time to thoroughly investigate the exam design, content, materials and procedures. On the whole, teachers had a positive attitude to the new examination and no course participant questioned the necessity of replacing the current school-leaving examination: there was general agreement that the changes were unavoidable so as to make the exam more prestigious, objective, reliable and compatible. Feedback from participants after trying out the sample examination tasks shows that some of the initial fears and doubts had to be reconsidered and modified on the basis of student results: not all 'teacher fears' were well-founded.

It was generally acknowledged that the planned examination would have a positive washback effect on classroom practices: teachers felt that the new '*érettségi'* appeared to test what one teaches, or should ideally teach in the communicative classroom. From this it follows that the teacher who is using a relatively modern course book, and is familiar with and consistently applies the methodological principles, will not have a problem in preparing his/her students for the new exam. The measured skills, the text and task types are similar to the ones communicative approach-based course books represent and communication-oriented teachers focus on.

This report on teachers' experience with the exam material and content would not be complete, however, if some of their queries were left unmentioned. Some participants who represented smaller grammar schools, or certain types of vocational schools with few English language classes per week, argued that their schools were not 'ready' for successfully preparing students for this demanding

examination. Furthermore, there were worries about the amount of extra work the new exam might mean for teachers and students. Some participants acknowledged that they definitely needed methodology updating and in-service courses in testing to feel safer about the new examination.

The most important message unanimously formulated by course participants was that such informative courses are very important. Also, there was a demand to be supplied with the precise and final description of the new exam (in terms of content, level definition, administration procedure and evaluation). As a course trainer I also agree that the sooner teachers are given information about the exam content, materials and procedures the more likely it is that they can prepare their students for a radically different examination.