Chapter 4

Pilot Test Booklet Design

Charles Alderson and Öveges Enikő

Terminology

In this volume we refer to test questions in various ways. The most detailed level of specificity is the ***item***: an item is that component of a task which is usually awarded one point if correct, and zero if incorrect. Items are usually grouped into ***tasks***, where a task has its own rubric, or instructions, and consists of between 4 and 15 items, depending on what skill is being tested. In the case of Writing or Speaking tasks, however, there are no individual items, as performance on a task is a complex, integrated activity, which is rated by humans using specially devised marking schemes and rating scales. Test tasks are combined into test ***booklets***, which are usually categorised by the major skill being tested, thus reading, listening, writing and so on.

Background

The reform of English examinations for Years 10 and 12 of Hungarian secondary schools is the object of a joint Hungarian-British Council project, the agreement for which was signed by The British Council, OKI Budapest and OKI Szeged in Spring 1998. The Year 10 examination in English was due to go live officially in 2002, and the Year 12 examination in 2004. By that time, according to the Project agreement, a bank of pre-tested items will have been created, which is intended to consist of 2000 items for the Basic exam (Year 10) and 2500 for each of the Intermediate and Advanced level exams at Year 12. It is highly likely that pilot examinations will be held in the one or two years prior to the formal first administration of the examinations, both to test the logistics of delivering and scoring these new examinations, and as a means of informing the public, and especially teachers and students, about the nature of the examinations and their likely demands and levels.

As described in Chapters 2 and 3, the project has developed Test Specifications, Guidelines for item writers, procedures for the production of tasks and items, procedures for the editing of such items, and means of paying for the production and editing of the items. The project has also developed procedures for the piloting of items, and piloting of suitable tasks took place in the period December 1998 to April, 1999.

The CITO Project

As part of the preparation for examination reform, a joint Hungarian-Dutch project, known as the CITO project, piloted and calibrated 200 multiple-choice items for both English Listening and English Reading, a total of 400 items. Each item was responded to by 240 students, and thus a spiralling design was needed to ensure adequate coverage of items, a design which was partially replicated in our own pilots. (The total CITO sample was 1200 students in Year 12.) The items existed in OKI and could be used as anchor items, or for comparative purposes, although they do not necessarily match the Specifications developed for the current project. As will be seen, we decided to use some of these calibrated items as anchor items in the design of our piloting. They are from item banks for the Dutch C level, considered to be at the Council of Europe A2 level (Waystage). Level C is the third level in 6 levels to be attained in Dutch secondary education. It is taught in lower secondary

education (C–stream) and lower vocational education (C–stream). The following table illustrates the relationship between Dutch and Council of Europe levels.

Table 4.1: Dutch and Council of Europe levels

Level A	=	A1 (Breakthrough)
Level B	=	A1/A2 (Breakthrough/Waystage)
Level C	=	A2 (Waystage)
Level D	=	B1 (Threshold)
Level E	=	B1/B2 (Threshold/ Vantage)
Level F	=	B2/C1 (Vantage/ Effective Operational proficiency)

The production of items/ tasks

Rates of pay for item writers had been discussed, and the complexity of the issue had been acknowledged. There was a lack of experience in Hungary in producing test items for three distinct levels, and no precedents existed for appropriate payment for the production of items for different skills and at different levels of difficulty. A survey of international practice (see Appendix I) revealed great disparities in international practice, and the need was recognised for the Hungarian Examination Reform Project to develop its own practice and precedents. It was also apparent that the teams needed to gain experience in writing items at the three levels, and for the amount of work involved in producing items for different levels and different skills to be established empirically. Furthermore, it was considered that the progression of difficulty across levels might usefully be examined by having item writers produce items for all three levels in the first instance, on the assumption that they would be more likely to differentiate the different levels if they had to produce items at all three levels rather than at only one level.

Rates of pay were agreed which were generous in the context of Hungarian practice, and which would motivate item writers to engage in what was likely to be quite hard work over the summer months, but not so generous as to be profligate or beyond the budget for the Project. In addition, an element of quality control was built into the payment system. Item writers were required to sign a contract agreeing to produce suitable tasks, according to the July Specifications and Guidelines, as follows: one task, consisting of several items, for each of the three levels of reading and listening, and for use of English at Advanced level only, plus one task for each of the three levels of writing and speaking, making 13 tasks in all. If tasks were submitted by the deadline (September 15th, 1998), item writers would receive an initial payment, and, if sufficient items met quality requirements, or were revised to meet such requirements, an additional payment would be made on satisfactory completion of the contract. Item writers were asked to keep a record of how long it took to produce items for different skills and at different levels, so that appropriate rates of payment could be worked out in the light of the experience.

Items submitted

Tasks were received in usable form from 14 item writers by the deadline. Some item writers submitted more than one version of a task, some had a task missing, but a total of 182 tasks was received in time for the Editing Committee to consider them. The Editing Committee classified tasks into three categories: 'Not usable', 'Usable, possibly after editing, for the piloting'; and 'To be discussed at the October workshop'. These latter items were chosen to illustrate particular problems with item writing, as a means of providing useful feedback to all item writers during the planned public moderation discussions at the October workshop (at Erdőtarcsa). For the
Listening tasks, there had been no opportunity to listen to the tapes in advance of or at the Editing Committee meeting, and so members agreed to identify possible candidates for

piloting or discussion by October 1st.

It must be noted that in the normal course of events, items will be produced by individuals specialising in particular skills and/or levels, and they will be discussed and modified by teams before being submitted to the Editing Committee. In addition, the deadlines for the submission of items will be more reasonable than was possible in summer 1998, and the number of items to be produced per individual will certainly be fewer. We were very grateful to item writers for the hard work that went into producing items under unfavourable conditions, especially during their summer vacation. The quality of what was produced in such circumstances was very impressive.

The numbers of items considered suitable for piloting by the Editing Commmittee was as follows:
(B = Basic, I = Intermediate, A = Advanced)

WRITING		READING		SPEAKING	
B	6 tasks	B	6 tasks	B	3 tasks
I	1 task	I	3 tasks	I	4 tasks
A	1 task	A	2 tasks	A	1 task

USE OF ENGLISH

98 items in 5 groups

The record by item writer was as follows. As a matter of policy all items are anonymised centrally after submission and tasks/ items can be identified by item writer registration number only.

ITEM WRITERS:

1 = 5 tasks	8 = 1 task
2 = 2 tasks	9 = 1 task
3 = 1 task	11 = 3 tasks
4 = 1 task	12 = 2 tasks
5 = 5 tasks	14 = 2 tasks
6 = 2 tasks	15 = 5 tasks
7 = 2 tasks	

It should be emphasised at this point that we did not expect all items to be perfect. We knew from experience that even items that appeared to be working well could have problems when tried out on real students. That is, after all, why items should be pre-tested as a matter of principle. But we were also aware that some of the items were quite likely to have problems, both in the productive tests of Writing and Speaking, and in the Written booklets, either because of the texts chosen, the quality of the recordings or the task or item design. However, it was important to show item writers how their items had fared in the real world, since we believe strongly that item writers learn best by seeing how well their items perform, and then analysing how they might be revised and improved in the future.

Piloting plan

For large scale piloting, we planned only to administer the Written exam (Listening, Reading and Use of English). Plans for the piloting of the Speaking and Writing tests were developed separately and the Writing and Speaking tests were administered in early December 1998, on a much smaller scale.

In order to gain stable estimates of item characteristics for the Written exam, it was essential that each item be responded to by at least 200 students. This enabled suitable statistics to be

calculated. Given the number of items ultimately required by the project, and the need for these to be calibrated on a common scale of ability/ difficulty, it was clear that a design with anchor items would be needed in the long term as well as the short term, and it was thus essential to experiment with such designs at an early stage in the project.

Since many tasks were involved even at this first stage of piloting, it was clear that a number of different test booklets was needed. It was also essential to administer these booklets in a fixed order where possible, both to avoid cheating within one class, and to ensure adequate sampling of students across booklets within each class. If one booklet were given to one class and a different one to another, the risk would exist that the sample responding to one booklet would be more or less able than that responding to another booklet. However, given the obvious logistical problems with Listening tests, any one booklet had to be administered to an intact class, and we had to hope that classes were sampled relatively randomly, in other words, some test booklets should not inadvertently be taken only by good students, and other booklets by weaker students.

As items were to be trialled at all three levels of difficulty, it was essential that test booklets contained items at different levels of difficulty, so that comparisons are possible across levels and anchor items. Booklets also had to be of equivalent length.

The tests were to be administered in the following order: Listening first, followed by a break of about 15-20 minutes. Then the Reading/ Use of English booklets. The time available for piloting was estimated as follows:

Listening: 30 minutes
Reading: 45 minutes
Use of English: 30 minutes

In the case of the Writing test, we created four booklets, each with two tasks of roughly equivalent difficulty. These booklets would be distributed in order, to reduce the possibility of cheating (thus, booklet 1 to student 1, booklet 2 to student 2, and so on). The Writing test was planned to take one double class period of 90 minutes.

In the case of Speaking, we created four different combinations of tasks, again of roughly equivalent difficulty, and we also included versions with instructions in English and instructions in Hungarian. We also experimented with administering tasks to single candidates talking to an interlocutor, and to pairs of candidates with minimal involvement from the Interlocutor.

Pilot Sample

Relatively representative (or at least not obviously biased) samples were needed of students in Years 10 and 12, balanced according to geographical distribution over the country, for large-city, small-town and rural schools, for grammar, vocational and combined schools, and for good, average and weak schools or classes. Details of the actual sample are given in the next chapter.

Since at least 200 students were required to respond to each item, and six different Written booklets could be constructed from available tasks and anchor items (in order to maximise items piloted), it was clear that a final sample of around 1000 students would be needed.

There were obvious problems in getting possibly Basic (or lower) students to take Advanced and Advanced students to take Basic items. However, there was no alternative but randomly to try out items of supposedly different difficulty levels on all students, to see how they scaled, since we could not know in advance, or even during the pilot administration, whether students were actually 'Basic' or 'Advanced', especially as these terms were simply undefined. Item writers were asked to classify their items according to the three levels, but it was of course highly likely that item writers were simply wrong about difficulty of items, and therefore we would have had no basis for selecting, say, 'Basic' items for use with potentially Basic students and vice versa.

For the productive tests of Writing and Speaking, we did not require large samples, or even

representative samples, since we were also experimenting with the logistics of delivery, the problems of marking, and so on. Cooperative schools were identified in Budapest, Pécs and Szeged for these pilots – see next chapter.

The pilot booklets
Taking into account all the above considerations, we drafted a number of different booklet designs, and finally agreed on the following. The numbers indicate the item writer registration code, and the level indicated is simply the level estimated by the item writer. This would be confirmed or disconfirmed as part of the analysis of results:

Writing

Booklet One
a) 1 Basic: Royal Mail Pen Pal form-filling
b) 15 Basic: Response to advert for cars in The European

Booklet Two
a) 1 Intermediate: Letter to join a Penpal club in reply to ad
b) 1 Advanced: Letter in response to Discover USA leaflet

Booklet Three
a) 3 Basic: Letter to parents, prompted by pictures
b) 4 Basic: Letter prompted by diary entries

Booklet Four
a) 11a Basic: Description of person based on picture of office
b) 5 Basic: Letter to host family re stay in UK

The rationale for these designs was that we would have tasks of roughly similar difficulty together in a booklet, so that students would not be faced with enormous disparities of difficulty levels. In any case, there were very few Intermediate or Advanced tasks that were considered suitable for piloting. We also wanted to ensure a spread of task types across booklets, insofar as we had a variety of task types available.

Speaking

Booklet One
Task One: 12 Intermediate: Poster selection
Task Two: 2 Basic: Restaurant role play
Task Three: 14 Intermediate: Hotel roleplay, round trip

Booklet Two
Task One: 11 Basic: Son Bob's trip
Task Two: 7 Intermediate: Compare pictures of rooms
Task Three: 12 Intermediate: Penpal plans

Booklet Three
Task One: 15 Advanced: Select and describe a picture
Task Two: 14 Intermediate: Compare two hotels
Task Three: 12 Intermediate: Penpal plans

Booklet Four
Task One: 1 Basic: Map task
Task Two: 14 Intermediate: Recipe task
Task Three: 14 Intermediate: Hotel roleplay, round trip

Again, the aim was to ensure as far as possible that students would not be faced with great disparities of task difficulty.

The Written Examination

Listening
Booklet One: CITO, 2 Basic, 15 Intermediate
Booklet Two: CITO, 2 Intermediate, 15 Advanced

'CITO' refers to the anchor task, of 10 items, taken from the CITO project referred to above. In fact, the anchor task we used was also used as the anchor task in that project. By using the anchor task, we would be able to compare the difficulty of the piloted tasks contained in different booklets, which would have been taken by different students, by calibrating all tasks against the anchor, or common, task.

Reading and Use of English
Booklet One: CITO, 12 Basic, 5 Intermediate, Use of English
Anchor [5a, 15a)], 2
Booklet Two: CITO, 1a Basic, 8 Intermediate, Use of English Anchor, 4

These two booklets would be administered to the same students who had taken a Listening Booklet, after a break. To reduce the possibility of cheating, and to ensure an even distribution of booklets across students, in any one class (all of whose students had necessarily taken the same Listening test) the Reading/ Use of English booklets would be distributed alternately, such that student 1 would get R/UE booklet One, student 2 would get R/UE booklet Two, student 3 would get R/UE booklet One and so on.

As we had more Reading and Use of English tasks to pilot than Listening tasks, and we only needed 200 or so students to take any item/ booklet, we decided to administer extra Reading/ Use of English tasks to students who had not taken any listening test. These were therefore longer booklets, since more time was available for administration as time was not required for a listening test. These were Reading and Use of English booklets Three and Four.

Reading and Use of English only:
Booklet Three: CITO, 6 Basic, 7 Basic, 2 Basic, 9 Intermediate, Use of English Anchor, 15b
Booklet Four: CITO, 5 Basic, 5 Advanced, 15 Intermediate, 11 Advanced, Use of English Anchor, 14 (sentence-based)

The detailed descriptions of these various booklets and the tasks they contain are found in Chapters 10 to 14, together with a discussion of the empirical results.

Questionnaires

In addition to the test booklets, we designed questionnaires to accompany the tests. In the case of the Writing and Speaking booklets, the questionnaires were brief and open-ended. They requested brief bio-data: name, age, class, school, and length of time learning English. After the tasks had been completed, students also filled out a brief open-ended questionnaire asking for their reactions to the tasks: attitude, difficulty/ ease, familiarity and whether they would like similar tasks in the School-leaving examination.

In the case of the Written tests, the questionnaires were more elaborate, and close-ended, as we had found the open-ended questions for the productive skills very difficult and time-consuming to process when we inspected them during the Manchester training.

An initial questionnaire gathered relevant bio-data: age, gender, type of school, year of study (9-12), number of years studying English, English lessons per week, number of years studying English outside school (eg.: private tutor), English lessons per week outside school, whether students intended to take the Érettségi, whether they intended to take the university

entrance exam, other foreign language learned at school, number of years studying other foreign language(s), number of lessons of this language per week this year, whether they had taken the State language exam – Advanced, Intermediate or Basic – and whether they had taken any international language exam.

At the end of the booklet, there was a second questionnaire asking about the difficulty level of each part of the booklet, and candidates' previous experience with each task. In addition, candidates were asked to write down their starting and finishing time on each task, so that we could calculate how much time was required by each task.

All questionnaires were in Hungarian.

Finalising the layout of the Written Examination booklets

After the tasks had been selected and put into the six booklets, we had to start working on their final layout. Making them user-friendly and using appropriate instruments of data collection are fundamental to get valid, reliable and assessable results.

Considerable care was devoted to creating the general layout of the booklets (arranging the longer tasks on two facing pages and the rest afterwards, and editing the text in a coherent format). Each test task was laid out in such a way that markers could enter marks in a form that was readily usable by data processors. The questionnaires were finalised by one of the data analysts, adding the necessary codes and layout that made them more manageable for data processing.

On the second page of the booklets a list of instructions in Hungarian was printed. This included guidance on what kind of pen to use, on the amount of time allotted to the test, and on where to write what. The students were warned not to use any dictionaries, not to forget to note when they started and finished a task or to complete the questionnaires enclosed. The instructions proved to be clear, as virtually all students responded appropriately. Samples of the final six booklets are contained in Appendix IV.

In the next chapter we describe how the sample population was identified and how the tests were administered.