Chapter 15

Piloting the Basic-Level Examination

Nikolov Marianne, Pércsich Richárd and Szabó Gábor

Background to study

As a result of the introduction of a new National Core Curriculum (NCC) in 1998, which was the most spectacular educational innovation in the 1990s, a new Basic-level examination was supposed to be introduced in 2002. This examination would be new to the Hungarian educational system, and was expected to be taken by all school-leavers at the age of 16. (Traditionally, most secondary-school leavers take the érettségi exam at the age of 18.) The status of the exam has, however, never been clear, as students going on to Years 11 and 12 have also been thought likely to volunteer to take the Basic examination in some subjects. (For more detail see Fekete et al, 1999.)

The background to the new exam is the restructuring of the state education system since 1989. Instead of the traditional 8 primary- and 4 secondary-school years, since the change of regime new types of schools have come to life on the 4+8; 6+6; 10+2 pattern, whereas the curriculum kept the 8+4 pattern. The development of the final version of the new NCC took six years and it promotes the 10+2 structure, traditionally alien to the Hungarian educational system. Since its introduction in 1998 when compulsory education ended at age 16, the government has raised the school-leaving age to 18, and initiated work on so-called 'frame curricula'. These innovatory curricula are supposed to bridge the gap between the NCC and the local curricula which all schools are supposed to have prepared. According to the most recent plans, the NCC will also cover the last two years (Years 11 and 12) of compulsory education, therefore it will need to be redesigned and extended. At the same time 'frame curricula' are to reflect changes in the NCC.

The above processes have been going on in a socio-educational context (see Nikolov 1999a) where much lip service has been paid to quality assurance in education, and most of the steps discussed above have been taken to demonstrate the action decision-makers intend in order to enforce their policy. In the meantime, the role of the national bodies responsible for the development of examinations has not been clarified. A new network of quality assurance has been established (OKÉV), but its role and its relationship with existing exam centres (e.g.: OKI Budapest responsible for developing the two-level School-leaving exam and OKI Szeged responsible for the Basic exam) have not been defined. Therefore, the background to the pilot project described and analysed below can be characterised as confused and professionally problematic.

Clearly, as a result, the status and future of the Basic exam has come under threat: the school-leaving age has been extended to age 18; the revision of the NCC is underway; frame curricula are being developed; Year 10 does not seem to represent a major milestone in the educational career of the majority of students. Despite all this, work on the new Basic exam has been going on in 1999 in all subject areas according to the original plan, since this was budgeted by the government. A major stage in the development of this exam was its first piloting in June 1999. This will be discussed in this chapter.

We will first outline the aims and rationale of the piloting, then describe those who took part. Thereafter, the schedule and administration of the June piloting will be detailed, followed by the results and a discussion of the classical and IRT analyses of sample

booklets and tasks. We will discuss the performance on the piloted tasks of students from different school types and years, with a variety of classroom hours of instruction. Finally, we will examine feedback from schools.

Aims and rationale

The aims and rationale of the June 1999 pilot project were as follows:

- The aim was to pilot tasks for all subjects in which tasks were ready to be piloted.
- In modern languages only English was to be piloted, as the rationale has been to base other languages on the English model at a later stage, depending on funding made available by the Ministry.
- Tasks in all subjects were to be piloted in a written format, therefore oral tasks were excluded.
- Listening comprehension tasks were also excluded, as the distribution of cassettes was not ensured.
- In all subjects tasks were administered in eight different booklets, therefore eight booklets had to be compiled for English as well.
- The allotted time was also predetermined: 45 minutes could be devoted to the English booklet.
- Besides piloting tasks, in the English pilot project we intended to explore some of the relationships between years of study, hours per week, task familiarity, other language proficiency exams and students' feedback on piloted tasks. Therefore, a questionnaire was integrated into the booklet.

Participants

Schools were encouraged to volunteer to participate in the piloting of new tasks. Therefore, the booklets were not piloted on a carefully designed representative sample of the target population, but on volunteers. This is why, despite the original aim according to which tasks were to be designed and prepiloted for Year 10, booklets were actually piloted in Years 8, 9 and 10. Schools were invited to indicate in what subjects they wished to pilot booklets.

The reason for this unusual recruitment procedure must have been due to the problematic background of the Basic exam. At the time of the pilot, school administrators already knew about the dubious status of the Year 10 exam, and it was feared, though never explicitly stated, that nobody would volunteer to trial the tasks. This liberal attitude on part of the OKI Szeged exam centre was probably meant to attract attention and volunteers to ensure the implementation of the pilot project.

From the total of 811 schools volunteering to participate in the pilot project, about 500 requested English tests. However, no data is available on overall school types across all subjects. Altogether 8212 students took the 8 different booklets for English as a foreign language, many more than would have been necessary for piloting the tasks reliably.

The schedule and administration of the June 1999 pilot project

The time schedule of the English pilot followed that of all subjects, as it was an integral part of the overall Basic pilot project. Despite the fact that several tasks had been designed during the first year of the English Examination Reform Project, in February

1998 there were simply not enough tasks to be piloted in eight booklets. Therefore, new Reading and Writing tasks had to be written according to the draft *Detailed Requirements* (Cseresznyés et al. forthcoming) in one month flat.

These new tasks were first trialled on family members, friends and colleagues before pre-piloting, and feedback from this experience was integrated into the tasks: rubrics were reworded, some texts were adapted, others rejected. Eight booklets, each containing two Reading and one Writing task, were then compiled. The pre-pilot was run in four secondary schools where about fifty Year 10 students took each booklet. The teachers participating in the pre-pilot took part on a voluntary basis, and as pay was not available, they were each given a dictionary as a reward.

In April and May 1999 data for all 24 tasks were entered into a computer, analyses were conducted and statistical tables produced. Some tasks needed editing: a few Reading items did not seem to work well, so they were omitted. Rubrics needed standardisation, and the layout also had to be modified. Two Writing tasks had to be reworded and the number of words required was standardised at 100 across all Writing tasks.

There was one worrying outcome related to the Writing tasks: in one of the schools about half of the students did not even attempt Part 3, though they had performed well on the Reading tasks. It turned out that, contrary to what we had asked in the letter accompanying the booklets, their teacher had discouraged students with less experience, claiming that the tasks had been designed for more advanced learners of English. Thus, students decided not to write anything, as nothing was at stake for them. In the other schools no such problems occurred. On the contrary, all students seemed to do their best.

As a follow-up, teachers were asked to compare the results of their students on the pre-piloted tasks to their typical performances. They received the statistical tables with all students' coded scores, so that only they could decode them locally. All teachers reported that good students tended to perform well, with less successful ones scoring lower. However, in two schools teachers did find a few examples where they expected their students to be less successful, so they were pleasantly surprised.

In the light of the statistical analyses, some of the booklets seemed to contain two easier Reading tasks, whereas others two more difficult ones; one of the easier ones was changed for a more difficult task to ensure that each booklet had a relatively easy and a more difficult Reading task. However, no attempt could be made to compile booklets of the same difficulty, as we had no access to any calibrated task. Also, text and task types were carefully checked in all eight booklets to avoid overlaps.

As a result of pre-piloting, one of the Writing tasks had to be excluded. It was a simple form filling task with an authentic form for people looking for an international pen pal. Whether it was to be considered more of a Reading task than a Writing task had already provoked some heated debate in the item writing team, but on the face of it the form seemed to be relevant and easy to fill in and then to score objectively. Also, form filling is a real life task, often the only Writing task people perform. Although the task was appropriate for the level and interest of the target age group, the pre-pilot revealed a serious flaw. When students got the booklets, their names were coded to avoid any bias in evaluation. However, in the Writing task they were asked to fill in their names, addresses and other personal data. Students seemed extremely sensitive to this contradiction and filled in weird data to conceal their identity. Some used names of famous stars, impossible exotic places for address, 99 or 101 for their age, and quite unexpectedly, a few put typical Roma (gypsy) names, indicating underlying issues beyond the scope of our study. We will never know if they would have behaved the same way in a live exam. Finally, the task was impossible to score and so it was dropped. As a result, this task type has been removed from the *Detailed Requirements*. Instead, a new task was designed based on diary entries, which seemed to work well.

The writing assessment scheme was trialled by five markers on seven tasks. After marking the papers with the help of the writing scale, they got together to edit the wording of some of the criteria in the light of the evaluation of the Writing tasks.

Finally, the pre-piloted, edited tasks were rearranged and edited for the eight final booklets to be piloted on a larger population in June. In May both the printed versions and the files were submitted to OKI Szeged to be distributed to schools.

In June all booklets were piloted in a variety of schools all over Hungary. Schools received master copies of the booklets by mail and they copied and administered them to their students without any central control on the day in June which they considered appropriate. Booklets were then mailed back to the OKI Szeged examination centre to be evaluated centrally. Schools were provided with keys for the Reading tasks and marking schemes for the Writing tasks, and teachers were encouraged to evaluate their students' performances locally.

In July and August 1999 the data was entered into a computer database, and a sample of the written tasks was marked, some of them double-marked, also centrally, in August and September. Some of the results were presented at the IATEFL Hungary conference (Nikolov 1999b), but no central effort has been made to make the results publicly available.

Results and discussion of the June 1999 pilot

The eight booklets, labelled A1, A2,…. D2, were taken by 8,212 students in total. Table 15.1 indicates the number of students completing each booklet.

Table 15.1: Numbers of students taking the eight booklets

Booklet	Number of students
A1	1142
A2	1025
B1	1155
B2	930
C1	1093
C2	909
D1	1059
D2	899

As Table 15.1 illustrates, the number of participants was extremely high. For the purpose of simply piloting tasks one or two hundred students would have been enough, but as more schools volunteered they were not turned down. It is also important to point out that from the above population only between one and two hundred students were from Year 10, the original target population.

Each booklet consisted of four A4 pages comprising two Reading tasks (P1+P2), one Writing task (P3) and a short questionnaire on students' background. All 16 Reading tasks were matching tasks of 7 or 8 items each. The top score on each booklet was 15 or 16 points on P1+P2, and 16 points on P3, a guided composition of about 100 words.

Characterisation of the piloted tasks

As has been pointed out, a total of 16 Reading tasks and 8 Writing tasks were piloted. Unfortunately, one of the Reading tasks (A2P2) was ruined in the final editing phase, as

the wordprocessor automatically merged the first two items, and as a result of this unexpected and unnoticed technical error, the task had only 7 items although there were 8 slots to be filled in. The lesson of this negative experience is clear: the final version of the booklets has to be checked very carefully to avoid such fatal errors. Time pressure cannot be an excuse.

The text types of the reading texts ranged from youth magazine articles (B1P1; A2P1; B2P1; D2P1) to science book texts (A1P1; B1P2; C1P1; C1P2; D1P1; B2P2; C2P1; D2P2) and newspaper articles (A1P2; D1P2; C2P2). All texts were authentic, and with a few exceptions meant for a wider age range of adience, they targeted teenagers. After pre-piloting, some of the texts were edited: for example the authentic name of the youth magazine *19* was changed to *She* to avoid confusion, or in tasks where a short text was to be matched with its title, the title was slightly edited to avoid simple word matching.

Despite all the efforts to design appropriate Reading tasks, some problems became evident. As members of the Editing Committee pointed out, some of the texts were shorter than desirable; quizzes and interviews were not listed among the text types for item writers; and some of the items encouraged lexical matching instead of reading. These critical points are valid and care needs to be taken to avoid such pitfalls in the future. Obviously, Reading tasks are hard to design, as a lot depends on the availability of suitable authentic texts of the required quality. Such texts are not easy to find. Despite these drawbacks, the statistical analyses of the piloted tasks are believed to reveal meaningful results.

As for the eight Writing tasks, they were of different types. One of them (A1P3) was based on a visual prompt, and involved more creativity than the other tasks. Students were required to write approximately 100 words about a woman's job, daily routine, etc. based on a picture of her office. This task seemed to elicit the most varied writing performances. The other seven tasks required letter writing in about 100 words. Two of them were based on diary entries (B1P3; D1P3), three were letters of introduction: one to a host family in England (C1P3), another to a new pen pal (C2P3), and the third to students on an exchange to stay with a Hungarian student's family (D2P3). In two tasks students were required to write a letter based on an authentic advertisement. As would-be visitors to England, in one they were to find out about an easy summer job (A2P3), in another to book accommodation at a Youth Hostel (B2P3).

Classical item analysis

Using classical item analysis, important features of the tasks were revealed. Because of the large quantity of data it is not possible to give an item-level description of tasks here, and discussion will be confined to the task level.

As Table 15.2 indicates, the difficulty of almost half of the Reading tasks was beyond the level of the proficiency of the population. Seven out of sixteen Reading tasks proved to be too difficult for the students in the sample – 'difficult' meaning the percentage of correct answers lower than 50%, a somewhat arbitrary definition, perhaps, yet a good indicator of performance in a sample of this size. Task A2P2 cannot be considered any further, as it was spoilt in the final editing phase.

Table 15.2: Test takers' performance on 16 Reading tasks

Booklet	Task	N	No. of Items	Mean Percentage	Mean	Std. Dev
A1	1	1142	8	67	5.3	2.5
A1	2	1142	8	35	2.8	2.4
A2	1	1025	7	75	5.2	2.1
A2*	2	1025	8	7	0.5	1.1
B1	1	1155	7	75	5.2	2.0
B1	2	1155	8	48	3.8	2.4
B2	1	930	7	76	5.3	2.1
B2	2	930	8	35	2.7	2.5
C1	1	1093	8	37	2.9	2.5
C1	2	1093	8	68	5.4	2.6
C2	1	909	8	43	3.4	2.6
C2	2	909	7	83	5.7	1.9
D1	1	1059	8	48	3.8	2.7
D1	2	1059	7	74	5.1	2.0
D2	1	899	7	65	4.5	1.9
D2	2	899	8	80	6.3	2.2

A primary aim of the booklet design was to pair tasks whose estimated level of difficulty matched the level of proficiency of the target population with tasks that were expected to be more challenging. Therefore, the difference between scores of tasks in each booklet was expected and is indeed gratifying, yet a more accurate level setting is needed to revise booklets, and some tasks clearly need to be dropped or used at higher levels of the examination.

The reliability indices of the tasks suggest that the task types used for testing reading skills were promising. As Table 15.3 shows, with numbers of items per task as low as 7 and 8 almost all tasks had a *Cronbach's Alpha* value of 0.8 or above – a value not often witnessed with item numbers lower than 15-20. (This was, however, likely to have been a very heterogeneous population, which would inflate the reliability figures.)

Table 15.3: Reliability and discrimination parameters of 16 Reading tasks

Booklet	Task	N	No. of Items	Cronbach's Alpha	Mean Biserial
A1	1	1142	8	.83	.89
A1	2	1142	8	.83	.90
A2	1	1025	7	.84	.98
A2*	2	1025	8	.72	.95
B1	1	1155	7	.80	.92
B1	2	1155	8	.83	.89
B2	1	930	7	.83	.96
B2	2	930	8	.82	.86
C1	1	1093	8	.83	.88
C1	2	1093	8	.87	.95
C2	1	909	8	.84	.88
C2	2	909	7	.86	.97
D1	1	1059	8	.86	.91
D1	2	1059	7	.82	.95
D2	1	899	7	.79	.82
D2	2	899	8	.87	.98

Mean biserial values (a measure of the task discriminating between students of low and high proficiency) also appear to be favourable: while a value of 0.8 is very acceptable, most values reach or go beyond 0.9.

IRT analysis

While the classical analysis of data reveals several important trends, an alternative approach to data analysis, namely Item Response Theory (IRT), seems appropriate in order to gain a more thorough understanding of how the items performed.

However, the complexity of IRT raises the question of what advantages it can offer as compared to classical procedures. As is commonly pointed out in the literature on testing, item difficulty in classical test theory is defined as the ratio of correct responses to total responses for a given item. Thus, the level of difficulty for any given item will always be dependent on the ability of the group it is administered to. IRT, however, makes it possible to define an objective item difficulty order, which can be matched with the person ability measures. What this means in practice is that items piloted on one group of examinees can be safely considered to be appropriate or unsuitable for a different but similar group. All this without repeated administrations of the items—a major practical advantage which classical test analysis could never provide. This also means that, by collecting items used in various different tests, we can set up an item bank, where items are stored and ordered according to their position on the difficulty continuum. Once again, the practical advantage of such an item bank is that once the mean ability of a given group is at least approximately determined, a complete test can be compiled using the pool of items without any further piloting.

Since the present project is obviously aimed at item bank construction, the application of IRT seems more than justified. Owing to various practical as well as theoretical concerns (see Wright and Stone 1979, McNamara 1996, and Pollitt, 1999), the Rasch model was found to be most suitable for this purpose.

A caveat concerning the analysis of the pilot data is in order, however. Although IRT makes it possible for items administered in different tests to be placed on the same difficulty continuum, this can only be done if some sort of link exists between the different tests. The

link can either be a set of common items or common persons in the test-taking population. In the case of the present pilot data, no such link exists between the different booklets. While the two tasks in each booklet are connected by the same population, the different booklets were taken by entirely different candidates and were made up of completely different items. Hence, a comparison of item difficulty figures across booklets is not possible. It is possible, however, to select tasks that functioned well in the piloting and use them as so-called 'anchor tasks' for future versions. In other words, these tasks could function as a common reference point for further piloting, thereby providing the link between different booklets.

Let us now take a closer look at how the tasks performed in the light of the IRT analysis. It needs to be made clear that, within the scope of this chapter, we do not have the space to present the analyses of all tasks and booklets. Instead, examples of typical performances will be presented.

The analysis of each booklet followed the same pattern. First, the data were analyzed with the BIGSTEPS program (Wright and Linacre 1992). Then, in a second stage, fit statistics for persons were examined, and misfitting persons were identified. The concept of person and item fit is fundamental to IRT. With the help of fit statistics it is possible to identify persons and items whose behaviour is inconsistent, thus misfitting the probabilistic framework of the measurement model. In the case of persons this means behaviour which is inconsistent with the underlying ability, resulting in numerous unlikely correct or incorrect responses. In terms of items, misfit can be interpreted as an indication of an item provoking a large number of unexpected responses from candidates.

As a third step, misfitting persons' responses were deleted from the data file. This was necessary, as inconsistent person behaviour affects the figures describing item performance. Thus, to get a realistic picture of how the items worked it is essential to guarantee person fit. It has to be noted here, however, that only a limited number of candidates can be rightfully regarded as misfitting. An unusually high rate would indicate gross item problems resulting in seemingly inconsistent person behaviour.

A second analysis was then carried out on the modified data set, resulting in the final analyses which include item difficulty estimates and item fit statistics for the identification of problematic items.

Let us first examine the results of the analysis of Booklet A1. With the help of IRT it is possible to compare directly the ability of candidates with the difficulty of items. This comparison is presented in Figure 15.1.

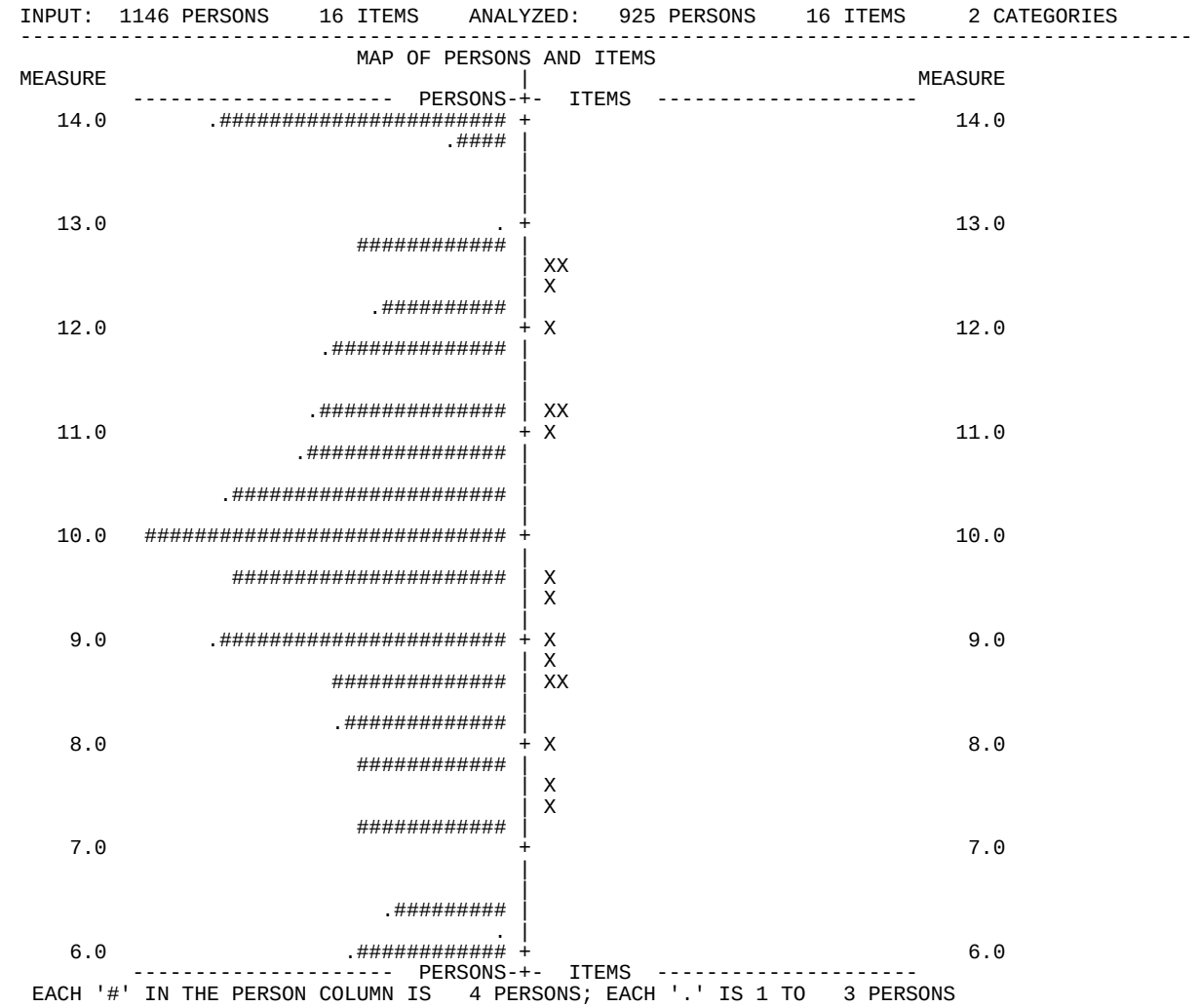

Figure 15.1: Map of persons' ability and items' difficulty in Booklet A1

On the left we can see candidates and their level of ability (MEASURE). On the right, items are plotted according to their difficulty along the same scale. As can be seen, candidates have an extremely wide range of ability. This is not surprising, as the sample included candidates of very different levels. Items also seem to cover a fairly wide range, but there appears to be a division between them. A fairly large gap can be observed between the difficulty measures of roughly 10 and 11, indicating that the items represent two more or less separate levels. It seems natural to assume that this division is to be explained by the difference in the two tasks included. This assumption can be verified by identifying the actual items and checking their difficulty measures. When we do so, we find that all of the items in the first task (A1P1) are located in the group of items with lower difficulty, while all but one items of the second task (A1P2) can be found among the high difficulty items. Apparently, these two tasks represent very different levels and should probably not be present in the same test. To find out which task then matches the level of the future Basic Exam, it is necessary to compare item difficulty measures with the person ability estimates of the Year 10 students in the sample.

The next stage is the checking of item fit statistics. As has been mentioned, through this analysis it is possible to identify items that provoke inconsistent responses. The type of fit that is most important for the analysis is technically called outfit, which indicates items that provoke unexpected responses from candidates whose ability is far from the difficulty of the item. A graphic representation of item outfit is presented in Figure 15.2

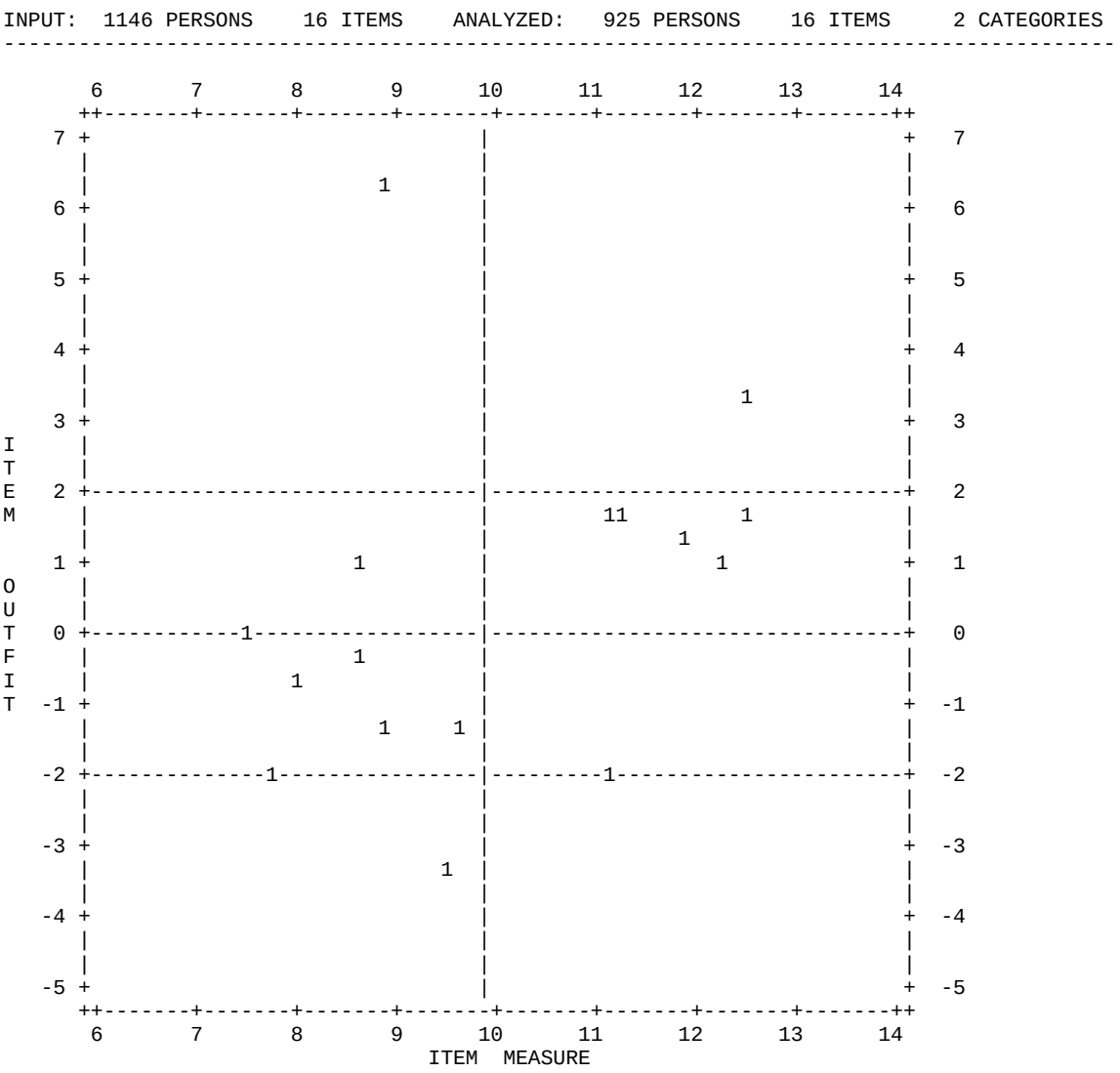

Figure 15.2: Map of item outfit in Booklet A1

The horizontal lines at 2 and –2 indicate the highest and lowest acceptable fit values. The vertical line at about ten indicates the mean of person ability. The items around and below –2, though not perfectly fitting, pose no particular problem, as negative fit values indicate so-called overfit. This means that an item functions in such a way that up to a certain level of ability no candidate gets it right; beyond that point, however, they all do. This kind of item performance misfits the model in the sense that a probabilistic approach necessitates at least some low-ability candidates getting a difficult item right or some high-ability ones getting it wrong. If this happens less frequently than the model would predict, the item is identified as a misfit (McNamara 1996:171). An overfitting item tends not to cause problems, so it is the truly misfitting items, those with outfit values higher than +2, that should be examined carefully. As can be seen in Figure 15.2, there are two such items in Booklet A1. Interestingly, one is part of the first task, and the other belongs to the second. It is at this point that a separate analysis of the tasks should be carried out to find out if this is a true representation of item fit.

After performing the separate analyses it was found that the same items appeared to be misfitting. While the nominal misfit values decreased somewhat, the conclusion remains the same: these two items are problematic in some sense and should be examined and revised, or possibly dropped.

At this point it is appropriate to consult traditional item statistics to find out if the reason for item misfit is a problem with item discrimination. Classical item statistics for Booklet A1 are presented in Table 15.4.

Table 15.4: Item analysis of Booklet A1

Item Statistics					Alternative Statistics					
Seq. No.	Scale -Item	Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	1-1	56	.89	.75	1	56	11	100	.75	*
					2	44	89	0	-.75	
2	1-2	79	.48	.61	1	79	52	100	.61	*
					2	21	48	0	-.61	
3	1-3	78	.56	.68	1	78	44	100	.68	*
					2	22	56	0	-.68	
4	1-4	54	.88	.73	1	54	12	100	.73	*
					2	46	88	0	-.73	
5	1-5	74	.60	.66	1	74	40	100	.66	*
					2	26	60	0	-.66	
6	1-6	63	.82	.74	1	63	18	100	.74	*
					2	37	82	0	-.74	
7	1-7	68	.72	.69	1	68	28	100	.69	*
					2	32	72	0	-.69	
8	1-8	65	.65	.58	1	65	35	100	.58	*
					2	35	65	0	-.58	
9	2-1	29	.66	.69	1	29	2	68	.69	*
					2	71	98	32	-.69	
10	2-2	38	.82	.71	1	38	3	85	.71	*
					2	62	97	15	-.71	
11	2-3	36	.85	.75	1	36	2	87	.75	*
					2	64	98	13	-.75	
12	2-4	24	.56	.66	1	24	3	59	.66	*
					2	76	97	41	-.66	
13	2-5	35	.77	.70	1	35	3	80	.70	*
					2	65	98	20	-.70	
14	2-6	24	.56	.65	1	24	3	58	.65	*
					2	76	98	42	-.65	
15	2-7	25	.62	.70	1	25	2	64	.70	*
					2	75	98	36	-.70	
16	2-8	69	.66	.59	1	69	33	98	.59	*
					2	31	68	2	-.59	

As can be observed, point biserial correlation figures, which represent item discrimination indices, show extremely high values. Although no strict rules exist concerning acceptable values of this figure, the lowest acceptable value tends to be defined around 0.3 (Alderson et al, 1995:82). The items that showed misfit in the IRT analysis are items number 8 (1-8) and 14 (2-6). Their discriminations, though lower than those of the majority, are still quite high. Consequently, the cause of misfit is to be found in some other property of these items. Probably these items measure a slightly different ability or sub-skill than the rest of the items.

A very different case is exemplified by the analysis of Booklet C1. Here, again, the first step is the comparison of person ability and item difficulty measures, as shown in Figure 15.3

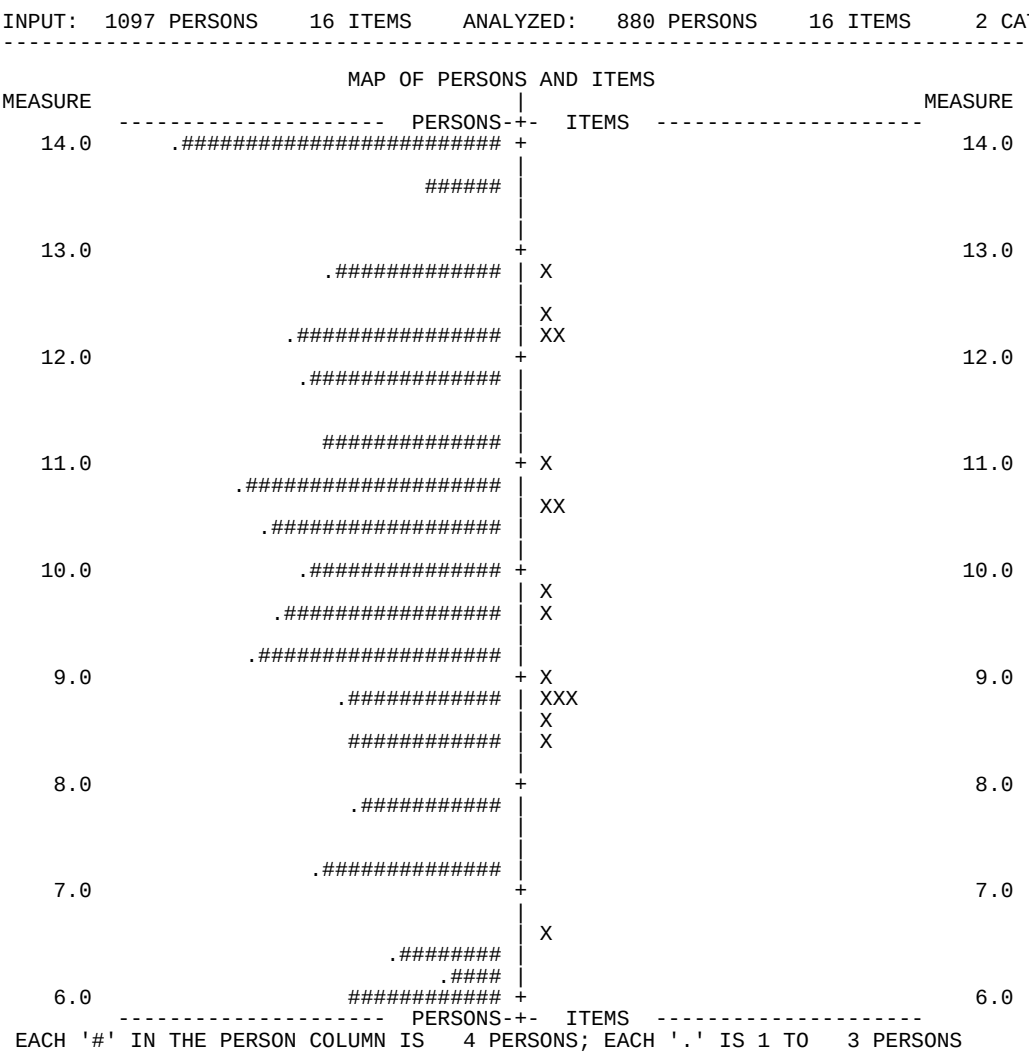

Figure 15.3: Map of persons' ability and items' difficulty in Booklet C1

Though the spread of candidate ability is wide in this case as well, Booklet C1 also shows a wider range of item difficulty. Also, the distribution of item difficulty seems more even, in the sense that no clear gap can be observed between two groups of items. Instead, three or four clusters of items can be identified. Interestingly, when the items are identified it can be observed that the top eight items in terms of difficulty all belong to task 1 in the Booklet (C1P1). The division is not obvious in Figure 15.3, yet it is clearer than in the previous case in the sense that all the items conform to the general pattern.

The most striking results, however, are yielded by the fit statistics. The map of item fit is presented in Figure 15.4.

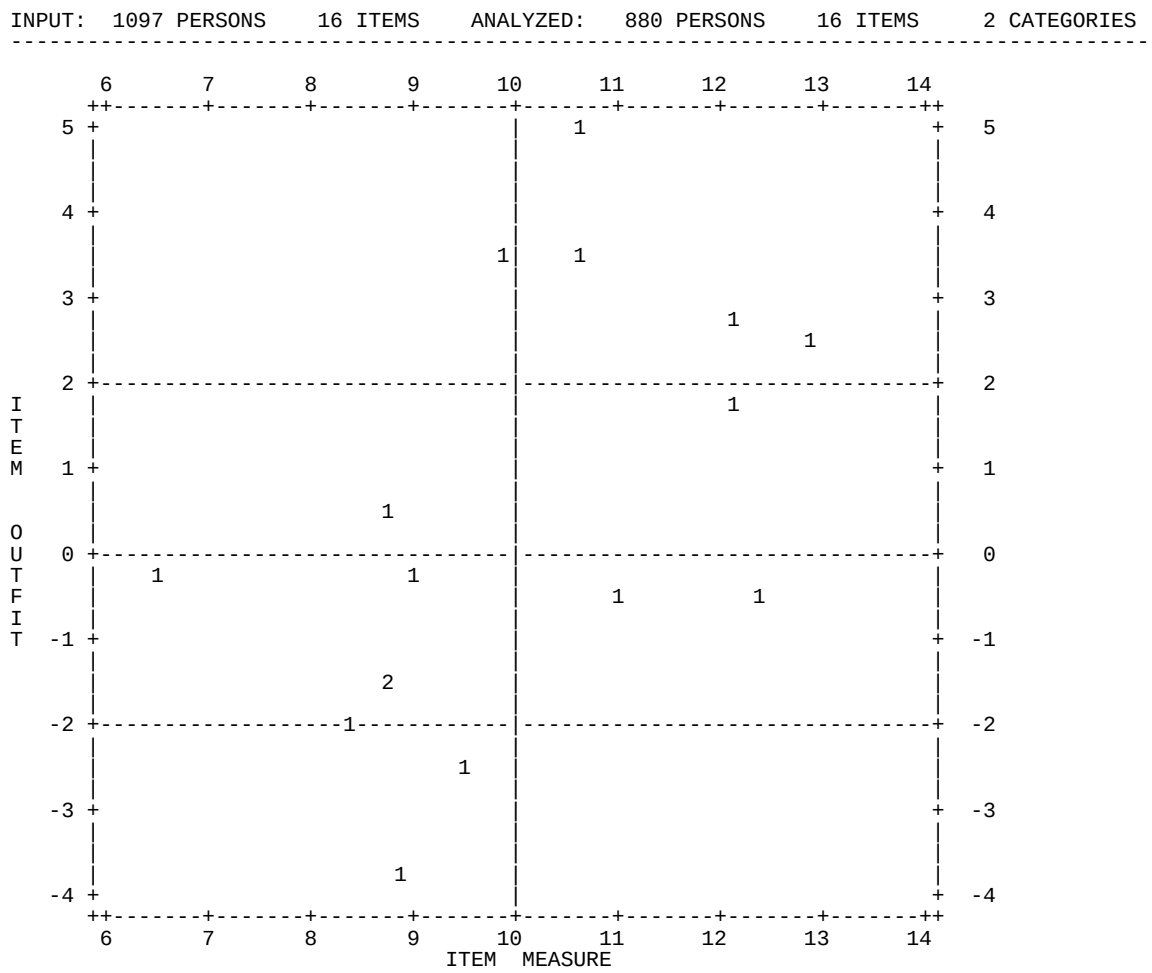

Figure 15.4: Map of item outfit in Booklet 5 (C1)

As can be seen, five items appear to be misfitting. This is a relatively large number, as the booklet only contained fifteen items in total. In fact, all the misfitting items belong to the same task: Task one. In other words, five out of eight items in this task misfit the model. If true, this means that Task one, though difficult, is useless, while Task two, though somewhat easier, includes no misfitting items, except for a few overfitting ones which cause no problems.

When a separate analysis of the two tasks is carried out, however, the results are dramatically different. This time only two items are identified as misfitting in the first task, and even these outfit values are barely beyond the acceptable limit (2.1 for both items). Moreover, one of these items (No. 4) was identified in the first analysis as a moderate, but definitely acceptable fit (outfit value: 1.6). At the same time, one misfitting item is detected in the second task with an outfit value of 2.6—a clear case of item misfit.

The indication of these figures appears to be fairly straightforward. The two tasks seem to work relatively well on their own, but they appear to be measuring very different things. In the course of the first analysis more items were in line with the construct defined by Task two (C1P2). These items included some items from the first task as well. Since the majority of the first task's items were measuring some other trait, they appeared to be misfitting. In the course of the second analysis, however, it turned out that the focus for the first task must have been different, and there appears to be only one item that is clearly problematic in the first task, the one that proved to be misfitting in both analyses (item No. 5).

Thus it is clear that only one of the two tasks could be used effectively for measuring either construct. Which construct is suitable for the purpose of the Basic Exam, however, can only be determined after a content analysis of the tasks themselves.

It seems clear from these examples, then, that IRT can add important information concerning item performance. Indeed, IRT-based item difficulty measures along with fit statistics are essential for item bank building. The next step should be the selection of a task that functioned appropriately in the light of both classical and IRT analyses, and use it as an anchor task. Then other tasks should be calibrated to this anchor by administering them together to yet another pilot population. This time the item difficulty measures would be fixed for the anchor items, and new item difficulty measures would be estimated for the other items in the other tasks. A similar procedure should also be applied in the case of brand new tasks.

One thing should be emphasized here, however. As in the example of Booklet A2, tasks may work well in the light of statistics but may measure something other than desired. Consequently, the selection of the anchor task has tremendous significance. As later tasks are calibrated to the anchor task in terms of difficulty and are compared to it in terms of model fit, it is of paramount importance that the anchor task represent the right construct.

Based on the statistics, there appear to be five tasks with characteristics which make them likely candidates to become anchor tasks. They are both tasks in Booklets A1 and C1 plus Task two in Booklet D1. All these tasks show favorable statistical characteristics in the light of both classical and IRT analyses. They all show high reliability and mean item discrimination figures, and they all have only a single misfitting item. As misfitting items need at least some revision, it does not seem to be appropriate to use the magnitude of misfit to order these items. According to classical statistics, Task C1P2 shows the best characteristics. Of the five tasks it has the highest reliability (0.87), and the highest mean discrimination figure (0.95). On the other hand, task D1P2, having the same mean discrimination and only slightly lower reliability, shows—in the IRT analysis—a wider range of item difficulty, distributed roughly evenly. Once again, the decision needs to be made on the basis of the subjective analysis of the tasks.

One more word of caution seems to be appropriate at this point. All the tasks examined in the course of the statistical analyses are matching tasks. The individual items in these tasks were scored and analysed separately. Indirectly, this assumes that a response to one item does not influence the answer to another. In IRT, this property of an item is referred to as the principle of local independence. In reality, however, it is quite likely that in a matching task items are not independent of one another. Hence, the reliability of the statistical figures is somewhat limited by the fact that the data do not meet the requirement of local independence, which underlines the importance of the subjective evaluation of tasks.

Text type of reading texts

Mean percentage scores of three major categories of text types (Table 15.5) show that students performed best on magazine articles, while some of the science book texts proved to be a great challenge. Since there were only three newspaper articles among the texts a valid comparison cannot be made in this respect.

Table 15.5: Mean percentage facility values across three types of text (values lower than 50% appear in bold)

Magazine articles:	*Science book texts:*		*Newspaper articles:*
B1P1: 75	A1P1: 67	**D1P1: 48**	**A1P2: 35**
A2P1: 75	**B1P2: 48**	**B2P2: 35**	D1P2: 74
B2P1: 76	**C1P1: 37**	**C2P1: 43**	C2P2: 83
D2P1: 65	C1P2: 68	D2P2: 83	

The findings correspond to the general perception of teachers and test specialists that texts on scientific topics tend to present problems for readers, yet test development seems to lack a proper linguistic theory for this.

Task types in Reading tasks

The task types of the 8 booklets can be categorised as follows:

Type 1: Match questions with given answers in dialogue
Type 2: Match science quiz questions with answers
Type 3: Match given text with missing parts to make it coherent
Type 4: Match title with short news

Table 15.6: Mean percentage facility values across three types of task (values lower than 50% appear in bold)

Task type 1	*Task type 2*	*Task type 3*	*Task type 4*
MAG.	SCB.	misc.:	NP.
B1P1: 75	A1P1: 67	**A1P2: 35 – NP.**	D1P2: 74
A2P1: 75	C1P2: 68	**B1P2: 48 – SCB.**	C2P2: 83
B2P1: 76	D2P2: 80	C1P1: 68 – SCB.	
D2P1: 65		**D1P1: 48 – SCB.**	
		B2P2: 35 – SCB.	
		C2P1: 43 – SCB.	

(MAG: magazine articles, SCB: science book texts, NP: newspaper articles)

The mean percentage scores of 4 task types (Table 15.6) suggest that task type 3 was the most difficult, but since five out of six texts came from science books, the cause of the low performance may well be the texts or topics themselves. More controlled research is needed to give an accurate explanation.

Background data – years of study and type of school

The tables below summarise the test results of Booklet A1 (n= 1142). All three tasks (P1, P2 and P3) are included to illustrate the main tendencies. The analysis is not extended to the remaining seven booklets due to lack of space but the main tendencies appear to be very similar throughout the rest of the booklets and tasks.

- ***How do years of study and performance on tasks compare to one another?***

Table 15.7: Average scores on three tasks and years of study – all school types included

Booklet A1	**n**		**mean**	**sd**		**mean**	**sd**		**mean**	**sd**
Years			**Part 1**			**Part 2**			**Part 3**	
1	57		3.19	2.27		1.33	1.45		4.37	5.72
2	98		4.93	2.51		2.13	2.02		7.32	8.36
3	71		5.34	2.47		2.76	2.49		11.90	10.51
4	110		3.95	2.74		1.70	1.76		5.80	7.73
5	218		5.18	2.42		2.66	2.42		9.14	9.37
6	177		5.60	2.23		2.63	2.40		11.90	9.36
7	135		6.02	2.23		3.16	2.49		13.76	10.32
8	158		6.09	2.32		3.85	2.68		15.39	9.99
9	42		7.10	1.63		4.55	2.38		19.05	8.41
10	29		6.69	1.82		4.72	2.74		17.21	8.48
11	6		8.00	0.00		5.33	1.97		23.00	4.73
no answer	40		5.10	2.51		2.43	2.32		8.63	9.22

Results appear to show a linear relationship between years of study and test scores although there are exceptions. However, the apparent differences are not statistically significant, and could have been the result of chance alone – the high standard deviations show that there was a very great deal of overlap in mean scores. Analyses of variance show that in most cases, years of study with a 1-year or even a 2-year gap are not significant.

In point of fact, only the contrasts of 3 and 4 years, 1 and 3 years and 4 and 6 years were significant, and even when there are gaps of three years, the only significant difference was between 4 and 7 years of study, whilst the contrasts 1-4 years, 2-5 years and 3-6 years showed no significant differences.

Moreover, the mean difference between 3 and 4 years of study is significant but in an unexpected direction (3 years of study results in higher mean scores than four years of study!). It is therefore logical to conclude that factors other than years of study must have an important role in candidates' performance.

- ***How do type of school and performance on tasks compare to one another?***

Among the three types of schools included in the sample, students of grammar schools (Type 2) appear to have performed best, followed by students in vocational (Type 3) and 8th grades of primary schools (Type 1).

Table 15.8: Mean scores across 4 booklets, school types compared

A1	n	mean	sd	min	max
Type					
1	511	12.51	7.82	0	32
2	369	16.91	8.33	0	32
3	244	11.25	7.24	0	31

A2	n	mean	sd	min	max
Type					
1	472	8.96	6.33	0	29
2	199	10.95	5.14	2	25
3	289	10.66	5.95	1	24

B1	n	mean	sd	min	max
Type					
1	512	13.06	8.20	0	31
2	376	18.15	7.99	0	31
3	244	13.47	6.73	0	31

B2	n	mean	sd	min	max
Type					
1	426	10.68	7.95	0	31
2	185	14.34	6.80	2	29
3	276	12.45	7.09	0	30

An analysis of variance of results on Booklet A1 indicates that there was a significant difference between the performance of 8th graders in primary schools and grammar schools, as well as grammar schools and vocational schools.

However, test-takers in vocational schools did not perform significantly better than test-takers in the 8th grade of primary schools, despite the fact that grammar and vocational schools were represented by 9-10 grade students.

- ***How does number of hours per week this year affect results on each task?***

While performance on the Reading tasks shows a relationship with the number of English lessons a week, only Part 3, a Writing task, shows a notable increase parallel with the number of lessons.

Table 15.9: Mean scores on three tasks, by lessons per week – all school types included

Booklet A1	n		mean	sd		mean	sd		mean	sd
Lessons per week			**Part 1**			**Part 2**			**Part 3**	
2	270		4.37	2.64		2.23	2.27		7.78	8.85
3	269		5.05	2.53		2.48	2.32		10.20	9.83
4	364		5.72	2.27		2.99	2.55		12.14	10.20
5	158		6.37	2.07		3.68	2.49		14.61	9.30
6	45		7.00	1.83		4.40	2.51		17.73	9.39
8	2		6.50	1.50		4.50	3.50		8.50	8.50
n.a.	32		4.97	2.78		1.81	2.17		7.16	9.02

Analyses of variance of the results of Booklet A1 indicate that the most obvious difference in performance was represented by the two ends of the spectrum: 2-5 and 3-6 lessons per week. Similarly, a 2-lesson per week difference also resulted in significant differences in test scores.

One-lesson per week differences tended to produce little difference in most tasks. However, it must be remembered that these figures cover a wide range of grades, from Grade 8 to Grade 12 in some cases. And the standard deviations are very high, indicating considerable variation even within one group.

- ***Did candidates planning to take the érettségi and / or entrance exam perform better than candidates not planning to do so?***

Table 15.10: Total scores and intention to take the 'érettségi' and an entrance exam to university – all school types included

érettségi	n	mean	sd		Entrance exam	n	mean	sd
no	297	11.20	8.04		no	544	12.34	7.94
yes	477	14.42	8.17		yes	130	17.07	8.03
hesitant	368	14.35	8.09		hesitant	468	13.99	8.30

Students who intended to take the érettségi and / or the entrance exam had a higher language ability (all differences were statistically significant). But a large number of students (368) were unsure whether they would take the érettségi yet performed just as well as those who intended to take it.

Task familiarity and performance on tasks in 2 booklets

In the accompanying questionnaire students were asked about their previous experience with the tasks. Results revealed that test takers who were familiar with the task type performed better. This supports the common sense expectation that teachers can prepare students for the actual exam by familiarising them with the various task types.

Table 15.11: Mean scores across 2 booklets, task experience compared

A1		n	mean	sd		n	mean	sd		n	mean	sd
		Pt1 – reading				Pt2 – reading				Pt3 – writing		
no		527	5.01	2.49		410	2.33	2.27		343	7.70	9.10
yes		349	5.93	2.35		478	3.20	2.54		535	13.82	9.95
dubious		149	5.06	2.64		137	2.31	2.31		147	7.99	8.95

A2		n	mean	sd		n	mean	sd		n	mean	sd
		Pt1 – reading				Pt2 – reading				Pt3 – writing		
no		408	4.94	2.27		477	0.56	1.12		393	5.12	7.72
yes		505	5.58	1.96		411	0.49	1.1		496	11.71	10.5
dubious		112	4.74	2.44		137	0.59	1.2		136	6.09	8.41

The difference in means between test-takers with and without previous experience (yes/no) was highly significant in the case of all three tasks.

Feedback from schools

Finally, we present a detailed qualitative analysis of the feedback schools sent to the exam centre together with the booklets filled in by their students. Note that few teachers took the trouble to put down their ideas. Out of about 500 participating schools only 17 sent in feedback: 15 from primary schools, one from a secondary school, and one failed to indicate the school type.

Among the 17 letters five contained only positive, six only negative, and a further six both types of remarks. Eight of the letters were explicitly apologetic for the low performance of the students and explained the perceived reasons.

Teachers made the following positive comments:

- The pilot was useful (tanulságos) (1);
- As for the Writing task, the marking scheme was good, it gave detailed information (árnyalt) (2);
- Tasks were good, of high quality (3);
- Students did the tasks with pleasure (szívesen);
- On the whole, the pilot was good, and hopefully it will be useful (3);
- Tasks were done in 45 minutes comfortably (1).

The list of critical remarks is longer:

- One task was problematic: 7 teachers pointed out that there were 7 items in task A2P2, whereas 8 in the key. Some rewrote the key, while others were more critical indicating that more care should be taken, "ha már hivatásszer•en ezzel foglalkoznak, legyenek szívesek komolyabb kontrolon átengedni a lapokat" [if you are doing this professionally, please take more care to control booklets more seriously.] Only one of the teachers identified the problem with the first item;
- Students should be allowed to use dictionaries (3);
- The booklets were of different difficulty (3);
- The vocabulary of the texts was too difficult (4);
- The Writing tasks were not primary-school tasks, they would have been more relevant for Year 10 students;
- Students have not had enough classes for this level: 2-3 classes/week are not enough (3);
- More variety of tasks would be necessary (2);
- The time was insufficient (1);
- Topics and vocabulary were not in harmony with what students had learnt: only what we teach must be assessed (2);
- *Project English* does not include such tasks, only on a lower level. (1)
- Accuracy was not emphasised enough: "Mindhárom feladat els•sorban a tanulók szókincsét, lexikai tudását méri, a nyelvhasználat helyessége háttérbe szorul. Általános iskolában a követelményrendszer nyelvtan centrikusabb, a szövegértést el•segít• feladatok száma kevesebb." [All three tasks assessed students' vocabulary, lexical knowledge; the accuracy of language usage is pushed to the background. The primary-school curriculum is more grammar-focused, the number of reading comprehension tasks is lower.] (1)

Other comments included the following points which could not be categorised either as favourable or critical:

- Our students could do these tasks because they were prepared for the Pitman Exam (1);
- Different tasks should be compiled for different school types (1);
- 100 words were not enough in the Writing task, 120-150 words would be needed (1);
- One teacher did not understand the marking scheme and asked: if someone got only 1 score on task achievement, could she get 8 on other areas?

To sum up, very few teachers commented on the pilot project. The vast majority were primary-school teachers, indicating the dubious status of the Basic-level examination:

secondary schools do not seem to be interested. This is in contrast with the traditional nationwide OKTV and OÁTV competitions, which tend to create considerable interest and result in more – and much more vehement and critical – feedback (Nikolov, Szabó and Kovács, 1996; Nikolov 1996).

Other reasons why there was a low response may include the following:

- As this was a pilot project, nothing was at stake either for students or for teachers;
- The timing of the pilot was extremely awkward, for June is the very end of the academic year, and both students and teachers were probably worn out;
- Schools were free to decide whether they marked the papers or not; they seemed to lack ownership over the project as most of them did not take the trouble;
- School administrators may have volunteered their students without asking their teachers, and this lack of involvement might have negatively influenced teachers' attitudes.

Despite the fact that teachers should have known that the pilot tasks were intended for Year 10 students, they complained about the level of the tasks. This may mean that some schools expected tasks designed for Year 8, or had had no access to essential information.

Although one of the tasks (A2P2) was clearly flawed, and students doing this task must have spotted the error, only seven teachers indicated this fact. Either the others did not notice, or they simply failed to point out the error.

The marking of the Writing tasks can also be considered to be a form of feedback from teachers. Few teachers bothered to evaluate their students' writings, and the vast majority of schools sent back unmarked booklets. As has been pointed out, teachers may have been tired at the end of the school year and they were not explicitly asked to do the marking, but they were offered the marking scale as an option.

Teachers seem to have an unclear picture of the rationale of the Basic Examination reform, as they do not seem to realize which year is to be involved in the new exam. This uncertainty has not been reduced or removed by any pronouncements from the Ministry. Secondly, some teachers insist on tasks being based on what they have taught, ignoring the variety of syllabuses and teaching materials used in Hungarian classrooms, and thus the impossibility of centrally set achievement tests. Finally, the lack of feedback might also indicate the low level of enthusiasm on the part of teachers; other studies have found them to be typically overworked and disillusioned (Nikolov 1999c).

Conclusion

In this chapter we have described the aims and rationale of the June 1999 pilot of the Basic Examination. We have characterised the participants, and described the administration of the tests. We believe that the analyses of the results of eight sample booklets have shown the strengths as well as the weaknesses of the tasks. We feel that data on how students performed from different school types after various years of instruction have revealed some important trends. Feedback from schools has also been important, as it indicates an area where more needs to be done. A great deal of effort has been put into the June Basic pilot. Readers can now decide for themselves how much experience has been gained at what price.