Chapter 13

The Reading Tests

Cseresznyés Mária

In the four *Reading* and *Use of English* Booklets, there are sixteen Reading tasks, including the Anchor task in each booklet, making a total of thirteen different Reading tasks. In addition to the Anchor task, Booklets 1 and 2 contain two tasks each, while both Booklet 3 and Booklet 4 contain four Reading tasks. The total number of items is 46 in Booklet 1, 58 in Booklet 2, 80 in Booklet 3, and 90 in Booklet 4.

General principles of task design

One of the main principles for item writers to follow was that tasks they designed had to conform to the *Specifications for the new school-leaving examination in English*. Apart from the Specifications for the exam, item writers were also required to consult the *Guidelines for Item Writers* document, which was intended to help them develop their items/tasks in accordance with the Specifications.

Generally speaking, the four sets of Reading tasks were intended to cover a range of different text and task types, with the aim of attempting to find appropriate measures of assessing different aspects of students' reading ability, as well as to give item writers feedback on how different task types performed. The tasks were meant to differ across the four sets also in terms of their difficulty level. We needed to obtain information, for one thing, on the actual levels of reading ability of Hungarian secondary school students, and, for another, on possible differences between levels of reading ability attained by students by the end of Year 10 and Year 12.

The tasks eventually selected for inclusion in the booklets were thought to be appropriate for assessing reading ability both in terms of different skills and strategies that are generally argued by researchers to be used by readers in different types of reading, and in terms of differences among levels of achievement.

In broad terms, both global and local comprehension and both macro- and micro (or enabling) skills were considered to be important. In accordance with this, some tasks in the four booklets focus on reading for gist, other tasks focus on students' ability to scan a text to locate specific information, while yet other tasks require students to read the text carefully for particular details.

The texts themselves were, with the exception of one text, taken from authentic sources. Authenticity of text was considered to be of crucial importance and, therefore, in designing the tasks, attempts were made to retain authentic features of the selected texts in terms of their content, the range and type of language used as well as, within the constraints of certain task types, in terms of their layout. The length of individual texts ranges between 200 words and 500 words, with the exception of one 32-word long text.

From the point of view of scoring, an important consideration was to develop tasks that could be scored objectively, or at least semi-objectively. Among the tasks included in the pilot booklets, the most commonly occuring task types are versions of multiple-matching, which is a form of multiple-choice. In fact, all tasks in the four booklets were of the selected-response type. Each correct response to an item scored one point, on all Reading tasks. The number of items per task ranges between 5 and 10.

The rubrics were given in the target language, that is, in English, using the format and wording produced by item writers for each particular task. Apart from the required type of response, item writers were recommended to include in the rubric some information on the type and/or topic of the text used, as well as on the number of answers candidates were expected to give in any particular task. In addition, an example answer had to be provided in each task.

Table 13.1 below gives a general overview of what is included in the four booklets (excluding the anchor test, taken from the CITO project as explained in Chapter 4). The booklets and tasks themselves are in Appendix IV.

Table 13.1: Overview of Reading Test tasks

Task No	Level	Skills tested	Topic and title of text	Words	Type/genre of text	Task type	Items
Booklet 1							
2	B	recognise cohesive ties in a text	Animals *It's Wild Penguins*	340	magazine article/ description	multiple-matching: clauses, phrases to gaps in text	10
3	I	Understand text structure	Travelling *Being wet got us a train ban ..*	460	magazine article/personal narrative	multiple-matching: paragraphs to gaps in text	7
Booklet 2							
2	B	find specific information, separate relevant from non-relevant	Animals No title	32	encyclopedia entry / process description	multiple-matching: pictures to text	5
3	I	select relevant information, recognise relationships between sentences of a text	The Queen's visit to hospital *Queen Meets Gardener Who Found Baby*	260	news report	sequencing: jumbled sentences of the summary of the article	9
Booklet 3							
2	B	Understand gist, separate relevant from non-relevant information	Advertising (car, furniture, holiday/travel, language learning, etc)	200	advertisement	multiple-matching: headings to short texts	8
3	B	find specific information by scanning, infer meaning	Films	260	film preview	multiple matching: questions to short texts	10
4	B	Understand gist, find specific information	Advertising (goods and services)	330	advertisement	multiple-matching: sentences to short texts	10
5	I	recognise cohesive devices, understand their functions across paragraphs	Art/the famous Portland Vase *A smashing case*	428	narrative from a book	sequencing: jumbled paragraphs of text	6

Booklet 4							
2	B	Understand relations between parts of a text	Lifestyle/Life of a TV personality *Fatherhood has transformed me*	370	magazine article	multiple-matching: paragraphs to gaps in text	5
3	A	Understand gist	Film, theatre	520	programme guide	multiple matching: sentences to short texts	8
4	I	Understand gist, find specific information, infer meaning not explicitly stated in a text	Pickpockets in Budapest *Police must act to stop thieves*	260	letter to the editor	multiple-choice questions	7
5	A	understand gist, recognise cohesive devices	Education, schools *Girls-only success 'based on selection'*	267	newspaper article	multiple-matching: words, phrases to gaps in text	10

Description of the content of the tasks and the results of analysis

Booklet 1

In Booklet 1, there are three Reading tasks. The first task, as in all four booklets, is the Anchor task described in Chapter 4.

Task Two

In the second task, students were required to read a 340-word long text presenting mainly factual information on the life of penguins. The task employs a multiple-matching technique in which clauses, and in some cases phrases, are removed from and placed after the text and students are to match the removed parts of sentences to gaps in the text.

The text, which provides factual information, uses transactional/ideational language and, apart from the use of contracted forms, it can be said to conform to the conventions of written language. From a different perspective, the text represents the language of description, which is reflected, above all, in the predominant use of present tense sentences, involving either stative verbs or intransitives (e.g., '*go*', '*fly*', or '*return*'). It employs fairly simple structures and, apart from a few topic-related words like '*flap*', '*moult*', or '*huddle*', basic vocabulary. It should be noted that this text was taken from the *Catch* magazine, which is a language teaching publication, and which cannot be said to be authentic, at least not in the strict sense of authenticity ('... what is authentic is what is not simplified and what is not pedagogic', Davies 1984: 185).

The task itself, containing ten items, was intended by the item writer to be a Basic level task, and was intended to measure students' ability to 'recognise cohesive ties in a text'. It is arranged on two pages, with the text on one page and questions on the other. The text itself is organised into four paragraphs, and is presented with its title '*It's Wild Penguins*' printed in bold. The rubric, apart from describing the task itself, includes information on the topic and type of the text, as well as on the required number of

answers. One answer is provided as an example. Students were to indicate their answers in boxes provided after the task.

Results

Descriptive statistics for the task are presented in Table 13.2, the facility values (F.V.), discrimination indices (D.I.) for each item are shown in Table 13.3. Table 13.3 also shows the IRT estimates of item difficulty (M).

Table 13.2: Descriptive statistics for Booklet 1, Task 2

N of Items	10
Mean	8.322
S.D.	2.316
Reliability	0.841
Mean % Corr	83
Mean Item-Tot Corr	0.656

Table 13.3: Facility value, discrimination index, and logit value of item difficulty for each item, Booklet 1, Task 2

Items	1	2	3	4	5	6	7	8	9	10
F.V.	66%	94%	93%	91%	94%	72%	80%	86%	81%	74%
D.I.	.65	.20	.22	.25	.20	.58	.37	.41	.33	.67
M	-.83	-3.25	-3.38	-3.08	-3.46	-1.21	-1.90	-2.41	-2.04	-1.42

As is clear from the figures above, this task was very easy for the population. The mean percentage score of 83 suggests that, despite the item writer's intention, it might be easy even for a Basic level task. The facility value of most items (7 out of 10) is above 80%, and in the case of four items, it is above 90% (Items 2, 3, 4, and 5). Items with such facility values do not usually discriminate well among students. Although, as can be seen from Table 13.3, of the seven easy items, the discrimination indices of three can be said to be acceptable (.37, .41, and .33 in the case of Items 7, 8, and 9), the discrimination indices of the items with the highest facility values (Items 2, 3, 4, and 5) are low (.20, .22, .25, and .20). There might be many different reasons for this.

As the IRT estimates of item difficulty show, Items 2, 3, 4, and 5, with logit values ranging between -3.08 and -3.46, are very easy items irrespective of the ability of students who have taken the test. The most important consideration is which factors influence the difficulty of an item. It may be the case that an item is easy and does not discriminate well because, like Item 2 in this task, it requires identification of one of two main elements of a parallel structure ('The emperor penguin is *the biggest*/and *strangest* of the penguins'). However, an item might be easy simply because it stands out among the other items in the task, which seems to be the case with Item 3 in this task. Item 3, as is clearly indicated in the text itself, is the only item which requires a whole sentence response, while, at the same time, among the options from which students can choose their answers, there is only one which contains a whole sentence. This makes the item very easy and results in failure of the item effectively to discriminate between students. This is, however, only one aspect of potential problems with the item. The other side of the coin is that an item may assess something that does not belong to the construct it aims to assess. Correct responses to Item 3 in this task may not tell us much about students' reading skills, their ability either 'to recognise cohesive ties in a text' or otherwise. Rather, they may be due to test-wiseness, as suggested above.

The figures in Table 13.3 suggest that, for our purposes, the best items in the task are Items 1, 6, and 10, which have acceptable facility values (66%, 72%, and 74%) and, at the same time, good discrimination indices (.65, .58, and .67). Lastly, on the basis of the IRT estimate of its difficulty reflected in the mean logit value of -2.298, this task as a whole is the third easiest of the seven originally intended by item writers to be Basic.

Students reported taking an average of 10 minutes to complete this task, ranging from a minimum of 3 minutes to a maximum of 32 minutes.

Task Three

The third task in Booklet 1 (the second without the anchor) is based on a 460-word long narrative. This task was intended to assess understanding of text structure, of how information is organised in a narrative. For this reason, whole paragraphs were removed from the text and placed in a jumbled order after it, and the task was to restore the narrative by matching the removed paragraphs to appropriate gaps in the text.

The text used in this task was taken from a teenage magazine and represents the most common type of narrative, the narrative of personal experience. The writer recounts her personal experience of one particular instance of travelling by train. It should be noted, however, that, in fact, the purpose of the writer, that is, the main function of the text, is to complain about certain aspects of train services that are closely related to her recent journey. In line with this, both the organisation of information and, accordingly, the particular variety of language involved demonstrates a mixed use of registers. At one level, this mixing of registers is manifested in the use of elements of more than one genre, including, apart from the main elements of narration, those of description and argumentation. However, it can also be observed at the level of linguistic features of the text, insofar as the text displays a mixed use of features of written and spoken/formal and informal language. Among these features, we find long, complex sentences with subordinate or embedded clause constructions, on the one hand, and contractions, certain discourse markers and vocabulary items which are more common in speech, on the other. To involve the reader as fully as possible in the narrative, the writer employs a range of evaluation devices, including, among others, the use of metaphors (e.g. '*it was like standing in a power shower*'), fixed expressions like '*One thing is for sure*', intensifiers of all sorts (e.g., *a huge clap of thunder*, *freezing cold*, *storming mad*), direct speech, etc.

The task itself contains seven items and was intended by the item writer to be an Intermediate level task. Similarly to the task discussed earlier, it is arranged on two pages, with the text on one page and the items on the other. In an attempt to retain authentic features of the text not only in terms of the range and type of language used, but also in terms of its layout, the text is presented in two colums with the title '*Being wet got us a train ban*' printed in bold. The rubric gives information about the type of text as well as on the required number of answers. An example answer is also provided.

Results

Descriptive statistics for the task are presented in Table 13.4, facility values, discrimination indices, along with the IRT estimates of item difficulty are shown in Table 13.5.

Table 13.4: Descriptive statistics for Booklet 1, Task 3

N of Items	7
Mean	3.783
S.D.	2.539

Reliability	0.865
Mean % Corr	54
Mean Item-Tot Corr	0.743

Table 13.5: Facility value, discrimination index, and logit value of item difficulty for each item, Booklet 1, Task 3

Items	1	2	3	4	5	6	7
F.V.	42%	53%	53%	42%	56%	74%	57%
D.I.	.82	.81	.54	.58	.84	.68	.85
M	.44	-.16	-.12	.51	-.30	-1.31	-.37

As can be seen from the figures above, this task, in accordance with our intention, was fairly difficult for the population. In any case, the mean percentage score of 54 implies that it was more difficult for the students than Task 2 (which had a mean score of 83%), resulting in a wider spread of scores (with a higher standard deviation) and, in general, in more effective discrimination. The mean logit value of -0.187 also clearly shows that there is a considerable difference between the difficulty level of this task and that of Task 2 discussed earlier (-2.298). This can be seen also from the figures (M) in Table 13.5, showing the difficulty level of each item.

Looking at the performance of individual items, it can be observed that the highest facility value (and, in fact, the only one above 60%) was generated by Item 6 (74%), with the rest of the items having facility values between 42% and 57%. The discrimination index of all seven items in the task is extremely good, but especially in the case of Items 1, 2, 5, and 7, where it is above .8.

On the whole, results of the analysis suggest that this task discriminates well among students with differing levels of reading ability. It should be emphasized, however, that it does so in terms of the differences between reading abilities of 'good' vs 'poor' readers, which does not necessarily coincide with the differences between the reading abilities of Year 10 and Year 12 students. From this point of view, it might be worth noting that, on this particular task, Year 10 students performed (even if only slightly) better than Year 12 students (mean facility values of 54.6 and 53.9). In fact, this difference between Years 10 and 12 is reflected in students' results on four of the seven items of the task. The respective items are Item 2 (55%/52%), Item 3 (55%/53%), Item 6 (77%/74%), and Item 7 (62%/55%).

Students reported taking an average of 11 minutes to complete this task, ranging from a minimum of 4 minutes to a maximum of 45 minutes.

Booklet 2

Task Two

The second task in Booklet 2 was designed by the item writer to 'test the ability to find specific information, separate relevant from non-relevant information'. For this, a short text containing 32 words was selected. It was taken from a (children's) encyclopedia and it describes how a frog develops from an egg. It consists of eight sentences and represents transactional/formal language, determined by the text type. All sentences of the text are short and simple, using the present simple tense to describe different stages of the process of the development of the frog. The vocabulary used in the text contains,

for the most part, specific topic-related words and phrases like, for example, *front legs*, *hind legs*, *lungs*, *wings develop, tail shrinks, tadpoles grow*, *fully formed frog*, etc.

The text is accompanied by six drawings which are meant to illustrate six stages of the development process. The task requires students to match the pictures, marked with letters, to individual sentences of the text, which are numbered. The required response to the task is thus of a selected type, in which students are to choose the appropriate number of the correct answer. The difficulty of the task is increased by the fact that, while there are six stages illustrated by pictures, there are eight sentences from which to choose those that describe the stages actually shown in the pictures. Of the six pictures, one is used to provide an example, which thus leaves five items for assessment. The task was intended by the item writer to be Basic, and is arranged on one page.

Results

Descriptive statistics are presented in Table 13.6, facility values, discrimination indices, and logit values of item difficulty are shown in Table 13.7.

Table 13.6: Descriptive statistics for Booklet 2, Task 2

N of Items	5
Mean	3.818
S.D.	1.385
Reliability	0.707
Mean % Corr	76
Mean Item-Tot Corr	0.671

Table 13.7: Facility value, discrimination index, and logit value for each item, Booklet 2, Task 2

Items	1	2	3	4	5
F.V.	72%	83%	74%	59%	94%
D.I.	.48	.44	.44	.67	.16
M	-1.15	-1.90	-1.47	-.46	-3.22

As is apparent from the mean for the task and the facility values of individual items, this task was relatively easy for students, which is not surprising in the light of the fact that it was intended to be a Basic level task. The fact that it was easy for students may at least partially explain why both the standard deviation (1.385) and the discrimination indices of all items, with the exception of Item 4, are relatively low. From this point of view, of all items in the task, Item 5 is the weakest item with a facility value of 94% and a discrimination index of .16. The logit value (M) of -3.22, especially when compared to the difficulty figures of the other items, also suggests potential problems with the item.

The reliability index (0.707) is not very high, which is probably due to the fact that the task includes only 5 items (one of which is very weak).

Note that on this task, as was the case with the previous task (the third task in Booklet 1), Year 10 students performed slightly better than Year 12 students. The mean facility value for the task is 76.2% in the case of Year 10, while it is only 75.6% in the case of Year 12 students. At item level, this difference is reflected in students' results on four of the five items in the task. These items are Item 2 (83%/82%), Item 3 (74%/73%), Item 4 (62%/57%), and Item 5 (95%/93%).

Students reported taking an average of 4 minutes to complete this task, ranging from a minimum of a few seconds to a maximum of 15 minutes.

Task Three

The third task in Booklet 2 is based on a 260-word long narrative taken from the Daily Telegraph. It was designed by the item writer 'to test the ability to select relevant information in order to perform a task, to recognise relationships between sentences of a text'.

The text itself, reporting on the Queen's visit to Kincardine Community Hospital in Stonehaven, represents the range and type of language common in news reports. It uses transactional, formal language, organised both at text level and the level of the sentence in accordance with conventions of the text type. In terms of textual organisation, this involves, among other things, the kind of reverse chronology typically used in news narratives, while at the more local levels, it involves the organisation of information in relatively long, complex sentences using embedded clause constructions and, also typically, passives.

For purposes of the task, a summary of the article had been written by the item writer, which, with its sentences mixed up, was presented after the text. The task proper was to arrange the jumbled sentences of the summary in the correct order, where the order was to show the chronological order of the events described in the original news report.

The task includes 9 items and is arranged on a single page. The text with its title *Queen Meets Gardener Who Found Baby* printed in bold is presented in two columns in the upper half of the page, while the items, put after the text, occupy the bottom half of the page. Response to the items, similarly to all the other Reading tasks on the pilot, is of the selected-response type, in which students are to choose the appropriate letter of the correct answer.

The rubric contains some reference to the text type as well as to the required number of answers; no information is given on the topic of the text.

With respect to its difficulty level, the task was meant to be Intermediate.

Results

Descriptive statistics are presented in Table 13.8, facility values (F.V.), discrimination indices (D.I.), and item difficulty figures (M) for each item are shown in Table 13.9.

Table 13.8: Descriptive statistics for Booklet 2, Task 3

N of Items	9
Mean	4.209
S.D.	3.335
Reliability	0.914
Mean % Corr	47
Mean Item-Tot Corr	0.767

Table 13.9: Facility values, discrimination indices, and estimates of item difficulty for each item, Booklet 2 Task 3

Items	1	2	3	4	5	6	7	8	9
F.V.	43%	49%	41%	33%	33%	42%	41%	62%	76%
D.I.	.79	.71	.75	.47	.53	.80	.79	.60	.48
M	.38	.07	.56	1.01	1.01	.45	.54	.62	-1.50

The mean figures for this task along with the facility values of individual items suggest that it was somewhat difficult for the students. The standard deviation (3.335), showing a

good spread of scores on the task, seems to imply the same. Looking at the performance of the task in the light of the results on individual items, we can see that most of the items (7 out of 9) have a facility value below 50%, and even of the remaining two, one has an F.V. slightly above 60% (Item 8 – 62%). While the facility value of most items can be said to be relatively low, their discrimination indices appear to be relatively high, which means they discriminated well among students. Even the lowest discrimination indices, generated by Items 4, 5, and 9 (.47, .53, and .48), are acceptable, but, as can be seen from Table 13.9, of the 9 items in the task, 5 have discrimination indices above .7.

The difficulty of the task can also be seen from the mean logit values of the items (M). As Table 13.9 shows, of the 9 items, 8 have a mean logit value above 0, and there is only one item where this figure is below 0, indicating that this item is easier than the rest of the items (Item 9, M: -1.50). Of the four tasks included in the first two booklets discussed so far, this task, with a mean logit value of +0.349, is the most difficult one.

Finally, there is one item in the task, Item 8, which appeared to be slightly more difficult for Year 12 than Year 10 students (F.V.: Y10/Y12 67% / 61%).

Students reported taking an average of 8 minutes to complete this task, ranging from a minimum of 2 minutes to a maximum of 25 minutes.

Booklet 3

Task Two

Task 2 in Booklet 3 was designed to test the ability 'to understand gist, separate relevant from non-relevant information'. The text forming the basis of this task consists of nine short texts, specifically, of classified advertisements. From each advertisement, a certain piece of information was taken out, in most cases in the form of a phrase referring either to the particular thing advertised in the text (e.g., *Tax Free Cars*, *London Tourist Flats*), or to the name of the advertising company itself (e.g., *Au Pair Bureau*). The task, employing a multiple-matching technique, requires students to match the removed phrases to gaps in the advertisements.

Given that, in terms of language use, authenticity of text has been retained, the language of each advertisement, by definition, displays features typical of the text type. Attempts were made to retain authenticity also in terms of layout, including typeface and size of type. The text, as a whole, uses a mixture of transactional and interactional language, with written and spoken features. While it provides, for the most part, factual information on different things advertised, and, to this extent, its language can be described as transactional (generally associated with written language), the organisation of language is clausal or phrasal rather than sentential (e.g., *Brand new cars direct from official importers*, *2 weeks minimum*, etc.), which is more typical of spoken than of written language. When complete sentences are used, they are generally short and simple, involving interactional language to express the writer's interest towards the reader (e.g., *'What sort of person are you?'*). The vocabulary used in the text is also relatively straightforward and simple.

The task includes 8 items and is arranged on two pages. The rubric contains information about the text type students are to read, but no information is provided with respect to the required number of answers.

The task was intended by the item writer to be a Basic level task.

Results

Table 13.10: Descriptive statistics for Booklet 3, Task 2

N of Items	9
Mean	7.189
S.D.	2.015
Reliability	0.748
Mean % Corr	80
Mean Item-Tot. Corr	.585

Table 13.11: Facility values, discrimination indices, and estimates of item difficulty for each item, Booklet 3, Task 2

Items	1	2	3	4	5	6	7	8
F.V.	82%	80%	69%	91%	94%	87%	61%	80%
D.I.	.38	.21	.63	.24	.17	.30	.48	.39
M	-2.51	-2.54	-1.70	-3.71	-3.76	-3.08	-1.28	-2.41

The mean score of 80% for the task and the similarly high facility values of individual items show that this task was very easy for this population. Of the 8 items, the facility values of 6 are between 80% and 94%. The two most difficult items are Item 3 with a facility value of 69% and Item 7 with a facility value of 61%.

On the basis of the mean logit value for the task of -2.624, this task is the easiest of all the tasks piloted. This is reflected in the logit values of individual items as well.

From the point of view of discrimination, the weakest items are Items 2, 4 and 5,, with a D.I. of .21, .24 and .17, while the discrimination indices of the remaining items are acceptable. The most discriminating items are Items 3 and 7, which are also the two most difficult items, with a DI of .63 and .48 respectively.

Students reported taking an average of 10 minutes to complete this task, ranging from a minimum of 1 minute to a maximum of 32 minutes.

Task Three

Task 3 in Booklet 3 was designed to test 'the ability to understand gist, find specific information by scanning, and infer meaning'. It is based on a 260-270-word long text, which, similarly to the text used in Task 2, consists of shorter pieces. In this task, however, students are required to read six film previews of about 40-50 words each. Given that the text type, along with the function of the text, is completely different from the one used in the previous task, the type and range of language involved is different. This particular text, consisting of film previews, displays, for the most part, features of written language. It uses both simple and more complex or more difficult language both in terms of linguistic structures and vocabulary.

The items included in the task consist of seven questions related to the films described in the previews. The task is to match each question to the description of the particular film in which the answer to the question can be found. Some of the items require more than one response, that is to say, in some cases, the answer to the question can be found in more than one of the previews (e.g., *Which film is about men and animals?*), in which case this is clearly indicated to candidates.

The task is arranged on two pages, with the text presented first, followed by the task/items. The rubric gives information on the text type students are to read, but not on the required number of answers. However, there are boxes provided after the task in

which students write their answers. In principle, the number of boxes should be taken as an indication of the required number of answers, but this may not have been clear to all candidates.

Concerning its difficulty level, the task was intended to be Basic.

Results

Descriptive statistics for the task are presented in Table 13.12, facility values, discrimination indices, and logit values for each item are shown in Table 13.13.

Table 13.12: Descriptive statistics for Booklet 3, Task 3

N of Items	10
Mean	7.403
S.D.	1.889
Reliability	0.628
Mean % Corr	74
Mean Item-Tot Corr	0.477

Table 13.13: Facility value, discrimination index and logit value for each item, Booklet 3, Task 3

Items	1	2	3	4	5	6	7	8	9	10
F.V.	75%	56%	87%	93%	93%	95%	85%	83%	46%	53%
D.I.	.32	.39	.33	.11	.23	.08	.28	.32	.50	.54
M	-2.02	-.97	-2.95	-3.77	-3.68	-4.18	-2.84	-2.58	-.47	-.82

This task, as expected, was easy for the population. This can be seen both by the mean percentage score of 74, and by the fact that, of the 10 items in the task, 7 have a facility value around 75%-80% or above. As the IRT estimates of item difficulty (M) also show, most of the items are very easy, with the logit value of seven of the items ranging between -2.02 and -4.18. The most difficult item in the task is Item 9, with a logit value of -.47. In the light of all this, it is not surprising that the discrimination of individual items is rather weak. The two highest discrimination indices were generated by Items 9 and 10 (.50 and .54).

With a mean logit value of -2.428, this task is at more or less the same difficulty level as Task 2 (Advertisements) discussed earlier.

Students reported taking an average of 9 minutes to complete this task, ranging from a minimum of 2 minutes to a maximum of 23 minutes.

Task Four

Task 4 in Booklet 3, similarly to the first two tasks discussed above, was intended to be a Basic level task. This time, the focus of the test is on the ability to find specific information by scanning. The task, like Task 2 discussed earlier, is based on a text consisting of classified advertisements. The total number of words in the 16 advertisements used in the task is approximately 330.

The text is fully authentic, displaying characteristic features of the text type. As is typical of the language of classified ads, the organisation of information is almost exclusively phrasal (e.g., *Live music for your wedding or social event*; *Specialist dealer with large stock of antique clocks*, etc.) The vocabulary used in the text includes words that represent simple, basic language (e.g., *perfect place*; *romantic meal*, etc.), but it also includes less frequently used, more specific, topic-related words like, for example, *bore hole drilling*, *water purification*, etc.

The arrangement of the texts on the page does not follow the usual layout of classified ads found in newspapers and magazines insofar as each advertisement is presented in one or two long lines stretching across the page. However, key words or headings in each ad (e.g. JOHN HAWORTH TELEVISION) use capital letters and are printed in bold.

The task employs a multiple-matching technique, in which students are to match statements about the interest of different people in the products and services advertised to the actual advertisements. While there are 16 advertisements for students to read, the number of items is only 10.

The rubric contains information about the text type and there is also some reference to the required number of answers.

Results

Descriptive statistics are presented in Table 13.14, facility values, discrimination indices, and the IRT estimates of item difficulty are shown in Table 13.15.

Table 13.14: Descriptive statistics for Booklet 3, Task 4

N of Items	10
Mean	6.794
S.D.	2.605
Reliability	0.814
Mean % Corr	68
Mean Item-Tot Corr	0.620

Table 13.15: Facility value, discrimination index, and logit value for each item, Booklet 3, Task 4

Items	1	2	3	4	5	6	7	8	9	10
F.V.	50%	85%	80%	69%	89%	80%	83%	52%	54%	62%
D.I.	.53	.47	.56	.77	.38	.53	.38	.59	.67	.56
M	-.69	-2.79	-2.41	-1.70	-3.26	-2.41	2.69	-.80	-.93	-1.32

As can be seen from the mean figures for the task, this task was slightly more difficult than either of the two other Basic tasks in the booklet. Of the three tasks, this one generated the widest spread of scores, with a standard deviation of 2.605. The mean logit value of the task as a whole is -1.91, which again implies that, of the three Basic tasks in the set, this task represents the highest difficulty level.

Concerning the performance of individual items, five items in the task have a facility value of 80% or above, while the F.V. of the rest of the items ranges between 50% and 69%. The most difficult item is Item 1, with a F.V. of 50% and a logit value of -.69. With the exception of Items 5 and 7, all items seem to discriminate well. Even in the case of these two items, the discrimination index can be said to be acceptable, given that it is, in each case, above .3.

Students reported taking an average of 12 minutes to complete this task, ranging from a minimum of 3 minutes to a maximum of 60 minutes.

Task Five

In Task 5, that is, the last Reading task included in Booklet 3, students read a 428-word long narrative taken from a book entitled '*The World's Greatest Trials*'. The text is an individual section in the book, represents a 'complete' narrative discourse and, thus, in terms of textual organisation, it displays all important features of the genre. It uses transactional, formal language and conforms, in every respect, to conventions of written

language. Information is organised in long, complex sentences, using a range of different structures, more often than not, with multiple embedding. Features of written language are reflected also in the use of structural/discourse markers as well as in the use of vocabulary, including the employment of less frequently occuring words and phrases (e.g., *an octagonal table*, *ignominious bow*, *habitual intemperance*, *miscreant*, *smashed to smithereens*, etc.) as well as, occasionally, that of technical vocabulary (e.g., *Wilful Damage Act*).

The task was intended to assess understanding of text structure, of cohesive devices and their functions across paragraphs. The method of assessment involves a sequencing task, in which students are required to arrange jumbled paragraphs of the text in order.

The task includes 6 items and is arranged on a single page. Following from the nature of the task, the layout of the text cannot be authentic. However, the text is presented with its title '*A smashing case*' printed in bold and its source is also provided at the bottom of the text. The first item is done for students as an example, which, at the same time, indicates the beginning of the story. The rubric informs students about the text type they are going to read, but there is no mention of the required number of answers.

The task was intended to be Intermediate.

Results

Descriptive statistics are presented in Table 13.16, facility values, discrimination indices, and logit values for each item are given in Table 13.17.

Table 13.16: Descriptive statistics for Booklet 3, Task 5

N of Items	6
Mean	1.214
S.D.	1.390
Reliability	0.615
Mean % Corr	20
Mean Item-Tot Corr	0.566

Table 13.17: Facility values, discrimination indices, and logit values for each item, Booklet 3, Task 5

Items	1	2	3	4	5	6
F.V.	25%	27%	22%	24%	21%	22%
D.I.	.56	.33	.24	.23	.26	.30
M	.17	.51	.91	.79	.94	.94

This task was beyond the abilities of most students taking the test. This is clearly shown by the extremely low mean percentage score of 20, as well as by the similarly low facility values of individual items. The highest F.V. is that of Item 2, and even this is only 27%. The IRT estimates of item difficulty (M) also suggest that the items in this task are much more difficult than the items in any of the three other tasks in the booklet. As can be seen from Table 13.17, the logit value of all items is above 0, while the logit values in the case of the other three tasks in this booklet are not only below zero, but the great majority of them are below (or around) -2.0, and some are even below -3.0.

Be that as it may, difficult items should not necessarily result in weak discrimination. However, the discrimination indices are generally low, with the sole exception of Item 1, which has a D.I. of .56.

The mean logit value of the task as a whole is +0.71, and this task is the most difficult of all the Reading tasks piloted.

Students reported taking an average of 11 minutes to complete this task, ranging from a minimum of 2 minutes to a maximum of 45 minutes.

Booklet 4

Task Two

Task 2 in Booklet 4 requires students to read an approximately 370-word long magazine article written about recent changes in the life of a TV personality. It was designed to test the ability to 'understand relations between parts of a text'. For this purpose, five paragraphs have been removed from the text and placed in a jumbled order after it. The task is to match the removed paragraphs to gaps in the text.

In terms of language use, authenticity of text has been retained and, thus, the text displays the range and type of language typically used in magazine articles of the kind. At the level of textual organisation this means that the basically narrative discourse contains a fair amount of description in different segments of the text. In fact, there seems to be more description than narration in the article. At the level of the sentence, the language of the text appears to be relatively straightforward, even though many of the sentences tend to be somewhat long, with quite a few of them using complex, in some cases difficult structures. The vocabulary can be said to be simple inasmuch as it involves, for the most part, simple, frequently occurring words and phrases.

The task includes five items and is arranged on two pages. The text is presented with its title '*Fatherhood has transformed me*' printed in bold. The rubric gives information about the type and topic of the text, as well as about the required number of answers. No example answer is provided.

The task was intended by the item writer to be a Basic level task.

Results

Descriptive statistics are presented in Table 13.18, facility values, discrimination indices, and logit values for each task are shown in Table 13.19.

Table 13.18: Descriptive statistics for Booklet 4, Task 2

N of Items	5
Mean	1.462
S.D.	1.547
Reliability	0.722
Mean % Corr	.29
Mean Item-Tot Corr	0.692

Table 13.19: Facility values, discrimination indices, and logit values for each item, Booklet 4, Task 2

Items	1	2	3	4	5
F.V.	33%	26%	39%	19%	29%
D.I.	.19	.23	.43	.28	.26
M	.03	.49	.30	.95	.20

The low mean of 29% for the task, along with the facility values of individual items, is an indication of the fact that this task was far more difficult for students than intended

by the item writer. The IRT estimate of the difficulty of the task is +0.394, with the logit values of individual items ranging between .03 and .95. In fact, this task is among the four most difficult tasks piloted. This task is clearly well beyond a Basic level.

As can be seen from Table 13.19, the easiest item for students taking the test was Item 3, with the very low F.V. of 39%, while the most difficult item in the task is Item 4, with a facility value of 19% and a logit value of .95. The discrimination of all items is weak. The most acceptable D.I. (.43) was for Item 3.

On one of the five items in the task, Year 10 students performed better than Year 12 students (Item 1 F.V. 36%/31%).

Students reported taking an average of 11 minutes to complete this task, ranging from a minimum of 2 minutes to a maximum of 56 minutes.

Task Three

Task 3 in Booklet 4 was intended to focus on gist reading. It is based on a text which represents the language of programme guides. The approximately 520-word long text consists of nine shorter texts providing brief accounts of different films and theatrical performances. For each film and theatrical play, a topic or summary sentence has been produced by the item writer, and the task is to match these summary sentences to the original descriptions.

In accordance with its main function, the text uses transactional language. Information is organised in long, complex sentences employing, for the most part, the kind of vocabulary characteristic of written discourse in general. Some specific, topic-related words and phrases, like, for example, *snarls*, *traitor*, *villains*, *local stables*, *emotional devastation*, etc, are likely to contribute to the difficulty of the text.

The task includes 8 items, and is arranged on two pages, with the items preceding the text. The text itself is presented in two columns.

The rubric gives information about the text type students are to read, and there is also some reference to the required number of answers. An example answer has been provided.

The task was intended by the item writer to be an Advanced level task.

Results

Descriptive statistics are presented in Table 13.20, facility values, discrimination indices, and IRT estimates of item difficulty are shown in Table 13.21.

Table 13.20: Descriptive statistics for Booklet 4, Task 3

N of Items	8
Mean	3.491
S.D.	2.186
Reliability	0.685
Mean % Corr	44
Mean Item-Tot Corr	0.559

Table 13.21: Facility values, discrimination indices, and logit values for each item, Booklet 4, Task 3

Items	1	2	3	4	5	6	7	8
F.V.	38%	32%	35%	41%	53%	57%	50%	44%
D.I.	.33	.23	.36	.44	.28	.42	.46	.51

M	-.23	.05	-.10	-.34	-.90	-1.14	-.82	-.56

As is clear from the figures shown in Table 13.20, this task, with a mean of 44%, was less difficult for students than Task 2 (intended to be Basic) in the same booklet. The mean logit value for the task is -0.505, which, however, suggests that the difficulty level of the task is definitely higher than that of most tasks intended to be Basic. Of all 12 Reading tasks piloted, with the above logit value, this task occupies the middle point on a continuum between the easiest and most difficult tasks. The logit value of individual items ranges between -1.14 and .05. The easiest item for students was Item 6, with a F.V. of 57% (and a logit value of -1.14), and the most difficult one was Item 2, with a F.V. of 32% (and a logit value of .05). Generally speaking, the items did not discriminate well, though it should be noted that, apart from Items 2 and 5, the discrimination indices are acceptable. The highest D.I. is .51 generated by Item 8.

On two of the eight items in the task, Year 10 students performed better than Year 12 students (Items 6 and 8 – facility values 61%/52% and 49%/42%).

Students reported taking an average of 14 minutes to complete this task, ranging from a minimum of 2 minutes to a maximum of 32 minutes.

Task Four

Task 4 in Booklet 4 was intended to assess the ability 'to understand gist, find specific information, and infer meaning not explicitly stated in a text'. It was meant to be an Intermediate level task.

The 260-word long text used in this task was taken from The Budapest Sun, an English language Hungarian newspaper. Concerning the type/genre of the text, it is a letter to the editor, which was written to draw people's attention to the dangers of pickpockets in Budapest and suggest possible ways of avoiding these dangers.

The text uses transactional language, involving features that characterise formal, written discourses. At text level, the organisation of language reflects a problem-solution pattern. In the first half of the text, the problem is described through an example presented in the form of a narrative, which by definition employs mainly narrative structures, while the other half of the text, embodying a response to the problem, displays the language of advice. Both in terms of linguistic structures and vocabulary, the language of the text can be said to be relatively straightforward and simple.

The task consists of seven 4-option multiple-choice questions and is arranged on two pages, with the text coming first. The text itself is presented with its title '*Police must act to stop thieves'* printed in bold, and is arranged in two columns. The author of the text is also indicated.

The rubric contains no information about either the type or topic of the text, and there is no indication of the required number of answers either. An example answer is provided.

Results

Descriptive statistics for the task are presented in Table 13.22, facility values, discrimination indices, and the mean logit values for individual items are shown in Table 13.23.

Table 13.22: Descriptive statistics for Booklet 4, Task 4

N of Items	7

Mean	3.406
S.D.	1.597
Reliability	0.516
Mean % Corr	49
Mean Item-Tot Corr	0.497

Table 13.23: Facility values, discrimination indices, and logit values for each item, Booklet 4, Task 4

Items	1	2	3	4	5	6	7
F.V.	45%	85%	38%	18%	30%	52%	73%
D.I.	.55	.08	.37	.14	.37	.38	.48
M	-.60	-2.73	-.21	1.02	.20	-.90	-1.91

This task is at roughly the same difficulty level as the previous task, that is, Task 3. This is shown both by the mean of 49%, and the mean logit value of -0.733 (cf. the figures 44% and -0.505 for Task 3). It seems, however, that, in terms of the difficulty of individual items, this task is less balanced than Task 3. Of the seven items included in this task, two were very easy for students, one with a F.V. of 85% (Item 2), the other with a F.V. of 73% (Item 7), while some other items, especially Item 4, with its F.V. of 18%, were extremely difficult for them.

The same unbalanced picture is shown by the IRT figures. As can be seen from Table 13.23, the easiest item (Item 2) has a logit value well below zero (-2.73), while the most difficult item (Item 4) has a logit value above zero (1.02). The discrimination indices of the items are generally rather low, but especially in the case of two items, namely, Item 2 (.08) and Item 4 (.14). From this point of view, the best item in the task is Item 1, with a D.I. of .55.

On two of the seven items in the task, Year 10 students performed better than Year 12 students (Items 3 and 6 – facility values 39%/36% and 56%/48%).

Lastly, with regard to the task as a whole, it should be noted that its reliability (0.516) is very low.

Students reported taking an average of 9 minutes to complete this task, ranging from a minimum of 1 minute to a maximum of 37 minutes.

Task Five

Task 5, the last Reading task in Booklet 4, is based on a 267-word long newspaper article taken from the *News* section of *The Times Educational Supplement*. It was intended by the item writer to assess 'understanding of gist, the ability to recognise cohesive devices in a text'.

The article, entitled '*Girls-only success 'based on selection*'', reports on findings of a recent investigation into the factors that may affect the success of single-sex versus co-educational schools. Following from its authentic nature, the text displays the range and type of language common to research reports of the kind published in newspapers, and is organised in accordance with conventions of this text type. At sentence level, information is typically organised in long, complex sentences, with the majority of them relying on multiple subordination or embedding of clauses. The use of passive structures, of both direct and indirect or reported speech, characterises the language of the text.

For the purposes of the task, the beginning of the first sentence in each paragraph was removed from the text and put after it, and the task was to match the removed parts of the sentences marked with letters to the numbered gaps in the text.

The task includes 10 items and is arranged on a single page. The arrangement of the text on the page does not follow the original two-column layout of the article. Its title, however, has been retained.

The rubric gives information about the text type students are to read, but not about the required number of answers. The task was meant to be Advanced.

Results

Descriptive statistics for the task are presented in Table 13.24, facility values, discrimination indices, and the IRT estimates of item difficulty for each item are shown in Table 13.25.

Table 13.24: Descriptive statistics for Booklet 4, Task 5

N of Items	10
Mean	2.645
S.D.	2.193
Reliability	0.709
Mean % Corr	26
Mean Item-Tot Corr	0.524

Table 13.25: Facility values, discrimination indices, and logit values for each item, Booklet 4, Task 5

Items	1	2	3	4	5	6	7	8	9	10
F.V.	11%	39%	14%	25%	57%	12%	12%	36%	31%	28%
D.I.	.13	.40	.14	.34	.49	.21	.18	.62	.45	.43
M	1.85	-.30	1.36	.55	-1.11	1.62	1.62	-.06	.15	.33

The figures shown in Table 13.24 suggest that this task was, as intended, difficult for the population. There are two points to make about the mean of 26%. One is that, although the task was, indeed, difficult for students, it was not more difficult than Task 5 in Booklet 3 which was originally intended to be Intermediate. Indeed this task was somewhat easier (the mean for Task 5 in Booklet 3 is 20%). However, recall that the two booklets were taken by different students, who may have had different levels of ability. Indeed, secondly, looking at the IRT analysis, it appears that these two tasks are roughly of the same level of difficulty. The mean logit value is +0.601 for this task, while it was +0.71 for the other task.

With respect to the performance of individual items, the easiest item in the task on the basis of both the F.V. and IRT figures is Item 5 (57% and -1.11), the most difficult one is Item 1 (11% and 1.85). The discrimination index of most items is acceptable. There are four weak items in the task, specifically Items 1, 3, 6, and 7, whose discrimination indices are below .3. The best discriminating item is Item 8, with a D.I. of .62.

On three of the ten items, Year 10 students performed better than Year 12 students (Items 2, 6, and 7 – facility values 41%/38%, 12%/10%, and 13%/11%). Item 4, with a F.V. of 25%, was of the same difficulty for both age groups.

Students reported taking an average of 10 minutes to complete this task, ranging from a minimum of 1 minute to a maximum of 33 minutes.

Conclusion

In this chapter, we have attempted to give a brief account of the content and results of the 12 Reading tasks piloted. With respect to the content of the tasks, we have, among other things, seen that

a) the 4 sets of tasks cover a range of different text and task types (see Table 13.1)
b) virtually all tasks are based on authentic texts (which, in view of the results, may imply that, contrary to the view held by many, the assessment of lower-level reading abilities does not necessarily require text simplification – simplifying, or adjusting the difficulty level of, the task itself may, to a greater extent, approximate real-life reading)
c) as the results of the analyses clearly show, the 12 tasks piloted differ, as intended, in terms of their difficulty level as well. The table below summarizes the results, focusing on task difficulty.

Table 13.26: Summary of difficulty of the 12 Reading tasks

Task Number	M	FV	B	I	A	
1	-2.624	80	X			B3
2	-2.428	74	X			B3
3	-2.298	83	X			B1
4	-1.910	68	X			B3
5	-1.640	76	X			B2
6	-0.733	49		X		B4
7	-0.505	44			X	B4
8	-0.187	54		X		B1
9	+0.349	47		X		B2
10	+0.394	29	X			B4
11	+0.601	26			X	B4
12	+0.710	20		X		B3

M – logit values, FV – facility values
B – Basic, I – Intermediate, A – Advanced
B1, B2, B3, B4 – Booklets 1-4

As can be seen, both the classical test analysis measure (F.V.) and the calibrated logit values of the tasks (M) show a wide range of difficulty. The facility values range from 20% to 83%, while the range of logit values is from +0.71 to -2.624. Viewed from this perspective, the selection of the 12 Reading tasks for the pilot can be said to be successful.

However, looking at the difficulty level of the tasks from the point of view of item writer intentions, we can say that, in some cases, there is a considerable discrepancy between the intended level of the task and the level implied by the actual results. Thus, for example, Task 7 with a F.V. of 44% and a logit value of -0.505 was originally intended to be Advanced, while Task 12 with a F.V. of 20% and a logit value of +0.71 was intended to be Intermediate. In this respect, perhaps the most striking difference can be observed in the case of Task 10, which, while originally intended to be Basic, is among the three most difficult Reading tasks piloted (F.V. 29%, M +0.394).

The results on the whole seem to suggest that the reading abilities of the test–taking population are very diverse. For a discussion of the issue of how this empirical data can be interpreted in terms of the two/three levels of the new examination, see Chapter 16 on standard setting.