Chapter 12

Listening Comprehension

Szabó Gábor and Nyirő Zsuzsanna

In this chapter we present an overview of issues related to the listening comprehension component of the new School-leaving examination. First, we discuss why it was considered essential to test listening tested separately in the new exam. Next, we describe the most important elements of the Specifications of the Listening Paper with special regard to text selection criteria and task types to be applied. The results of piloting are described next, and finally an account of issues to be resolved is presented.

The need for the Listening Paper

The current school-leaving examination does not test listening comprehension directly, but deciding to incorporate a Listening test into the proposed new examination was easy. Listening is one of the four basic language use skills and is widely recognized as important both in language teaching and testing. Consequently, it is only natural to include a listening component in any test battery intending to measure overall foreign language proficiency.

The Baseline Study showed that listening is a relatively neglected skill in the teaching process (Fekete et al 1999:243), despite the primacy of listening in language use. Yet including a Listening Paper seems absolutely indispensable in an examination where the communicative approach is emphasised. As a receptive skill, listening functions as a major tool for getting input for language learning, and, indeed, in real-life situations language learners tend to rely heavily on their listening skills (e.g. in getting information from the media, at railway stations, airports etc.). From a communicative perspective, then, the measuring of listening skills seems to be a basic requirement.

It is possible to argue that listening is measured indirectly in the course of the current oral test. However, face-to-face oral testing does not lend itself to task types reflecting various real-life listening tasks, and judgements in that test are, of course, subjective. We therefore decided to include a separate Listening Paper, which would also be more reliable.

Another reason why a separate Listening Paper is to be incorporated is to achieve beneficial washback. Teachers often tend to concentrate on exam requirements when deciding what and how they teach. In fact, it is probably the lack of listening comprehension in the present examination system that can, at least in part, be blamed for the low priority given to listening in many language classes. The new exam intends to facilitate the teaching of listening comprehension by setting a listening-related requirement that all school-leavers would have to meet.

Specifications

The requirements associated with this component of the exam are numerous and tend to be similar to requirements in many international language testing systems. Requirements include, for instance, the ability to understand the gist of a text, to understand specific information included in a text or to listen for implications.

The new School-leaving examination, however, is planned to have a two-level structure, which implies that students taking the Advanced level version may have to meet extra requirements. Indeed, the original version of the Specifications included special requirements for the Advanced level, such as the ability to follow complex and abstract chains of thought or to recognize ambiguities. Since then, however, we have proposed the same requirements, text types and task types for both levels. Thus, concerning requirements, we would leave out 'recognising ambiguities' and would have only 'recognizing and interpreting stylistic characteristics' and 'evaluating utterances and drawing inferences'. The difference between the levels would therefore be between the texts and not the tasks. At Advanced level the texts are longer and they contain more complex structures, quicker rhythm of speech, less common lexical items. (Of course, if further piloting proves that some tasks are not suitable for the Intermediate level, we will reconsider our suggestions.)

To achieve standardization, the whole Listening Paper is to be pre-recorded, including the instructions, which are to be in the target language, and which are to appear in print as well on the test sheets. The number of times a text is played is two for most tasks, but at Advanced level for gist listening tasks we originally proposed to allow a single listening, and this was indeed implemented in the pilot examination (Booklet Two, Task Three). However, in the light of the results of the piloting, and of consultations with teachers, especially on the in-service courses (see Chapter 18) and in view of recent developments in testing listening this option has been rejected (e.g. even in the latest of the Cambridge exams, the 'Certificate in Communicative Skills in English', all texts are heard twice).

Each Listening Paper is made up of at least three or four texts, depending on level. The length of the texts is about 2 minutes at Intermediate level and about 3 minutes at Advanced level.

The purpose of the Listening Paper is to assess how well candidates are able to understand the speech of speakers of the target language. We need to clarify what kind of speech and what kind of speakers are involved in this process. One of the most significant issues concerning a listening comprehension test is text selection, and so in the next section we examine some of the most important criteria related to this point, and discuss the issue of authenticity.

Issues in text selection

Since the new school-leaving exam is to be communicative, the issue of authenticity is important. Generally, texts tend to be regarded as authentic if they are produced by native speakers for purposes other than language teaching. While the latter part of this quasi-definition seems obvious enough, the presence of native speakers may be questionable. Since most speakers of English in the world are non-native speakers, and as they mainly use English for communication, it seems logical to raise the issue of whether including texts produced by non-native speakers is legitimate in a Hungarian school-leaving exam.

While counter arguments, including potentially negative washback, the countless varieties of non-native accents, or possible face validity problems can be cited, the first version of the Specifications allowed for experimenting with texts produced by proficient non-native speakers as well. This meant that we could investigate whether the potential problems actually materialized. However, this option was felt to be controversial and was subsequently rejected. The present modified Specifications do not allow experimenting with texts produced by proficient non-native speakers.

Besides non-native English, however, the problem of native varieties presents a dilemma, too. As one cannot talk about any variety of English being ‘standard,’ from a sociolinguistic perspective any native variety should have the same status. Certain varieties, however, are clearly more commonly encountered in the Hungarian setting than others, which means that testing proficiency through rarely–heard varieties may well be criticised. The present version of the Specifications does not identify any varieties as preferred, but the accessibility of authentic materials for test construction may well be a natural limitation and a solution to this problem.

Authenticity, however, is not solely defined by the speakers. As we have said, the content itself needs to be authentic. This means in part that the text should come from an authentic situation, but in a broader sense this notion could be interpreted to include the quality of the text’s topic being appropriate for the target population. In real life people listen to things that interest them, or things they need to understand for some special purpose. For instance, an announcement at an airport is not particularly ‘interesting,’ unless one is waiting in the departure lounge to board a specific flight.

In accordance with this, the Specifications require item writers to select texts on these grounds. Hence, text types for both levels include, among others, media announcements, directions, and interviews, conversations including two or more speakers and lectures on topics of public interest. Texts at Advanced level are more complex and may be characterized by a quicker rhythm of speech and/or less common lexical items.

The actual topics are to reflect the experience of the target population. Distressing topics, however, are to be avoided. Unfortunately, a topic that may seem particularly interesting for some, may turn out to be distressing for others. Thus special attention is paid to avoid using texts whose topic may cause concern in this sense.

The quality of the recordings is another crucial issue. Texts of poor recording quality cannot be used for measurement purposes. Arguably, in real-life communicative situations the conditions for listening may not be optimal. There may be too much background noise, the telephone line may not be clear, or the radio may not be tuned well. Consequently, one might consider the possibility of using recordings of such texts as well. Yet in an exam situation, which is stressful enough in itself, it would probably be inappropriate to further burden candidates with the added difficulty of text quality problems. Hence only recordings of high quality are to be used, and future plans include digitized recordings as well, which would guarantee CD-equivalent sound quality. It should be added here, however, that background noise which does not interfere with the actual text is acceptable, as it contributes to the creation of authentic sounding.

Apart from text selection, the other major element of the Specifications is the tasks themselves. The next section discusses this topic.

Task types

The task types described by the Specifications include a number of different techniques. Some of these, for instance selecting multiple choice statements or giving short answers, are to be applied at both levels, while others, such as choosing the best ending or the most appropriate summary of a text are to be employed mostly at Advanced level. Various types of sequencing tasks are also part of the list, and these are commonly considered to be somewhat problematic in terms of scoring. Traditionally, the scoring of such a task involves giving a point for each element of the sequence if it is placed correctly. This means, however, that if the first item in the sequence is misplaced, regardless of the rest of the sequence, no score can be awarded. Moreover, if the items are sequenced at random, and some items happen to be placed correctly, they earn scores, even though the candidate has obviously not understood the text. A new scoring algorithm, however, offers a solution to this problem by awarding scores for partially correct sequences by examining the relationship between the elements of the sequence (Alderson, Pércsich and Szabó, 2000).

The authenticity of the tasks themselves is also to be taken into account. In other words, where possible, a task should model some real-life situation in which the mechanics of problem solving are similar to the workings of the task. For instance, when one tries to identify the route to a particular location on a map based on auditory information, the mechanics of this activity will be nearly identical to those of an information transfer task requiring candidates to follow a route on a map as described by the text. Obviously, certain task types are more suitable for such 'authentication' than others.

Having described some of the most important points of the Specifications, we now examine the first piloting of the Listening Paper.

The piloting process

The first piloting of the Listening Paper included two booklets with three tasks each. See Appendix IV for copies of the tasks. One of the tasks was common to the two booklets, which made it possible to compare the performance of all the tasks and items through a Rasch analysis of the results. This common task (known as an anchor task) was multiple-choice, while the other four tasks included short answers, and marking on a map. Apart from the anchor task, one of the texts was used in both booklets, accompanied, however, by two different tasks. The purpose of this was to determine whether a difference in task design on the same input text would change the difficulty of the task significantly.

The piloting involved 513 candidates altogether, 244 of whom took Booklet 1, and 269 took Booklet 2. As described in Chapter 4, the candidates were high school students studying in various types of Hungarian high schools in different parts of the country. Their teachers volunteered to participate in the piloting and received book donations in return for their cooperation. The students themselves were requested to make every effort to solve the tasks. It has to be noted, however, that since the students received no grades for their performance, it is impossible to tell whether they took the tasks seriously. Hence, response validity is potentially questionable. This, however, is the constant problem of any piloting, which cannot be solved with a hundred percent certainty.

In the course of the actual piloting the participating students took the test under conditions that were envisaged to be similar to those of the would-be live exam. Invigilators included the students' own teachers and outside test administrators, who were provided with detailed instructions on how to conduct the pilot session. In the course of the piloting, the administrators also provided informal feedback on issues arising on the spot.

The data were analyzed using both classical and IRT (Item Response Theory)-based methods. In the light of the results, it seems clear that the tasks performed with varying degrees of success, as we discuss below. Most important, perhaps, is the fact that the levels associated with some tasks need revision. Some of the tasks considered relatively easy—and thus suitable for the Intermediate level of the exam—were found to be unexpectedly difficult, and some tasks originally intended for Advanced level proved suitable for lower level learners. The conclusion, obviously, is that level setting needs more time and more piloting. It may be the case, however, that task design itself may make all the difference between tasks involving the same text, and if the next piloting supports this, we will have to reconsider the issue and discuss the possibility of distinguishing the two levels according to task type as well as according to texts.

Statistical results

Table 12.1: The piloted Listening tasks: an overview

Task ID	Intended level	Input text type	Task type
Booklets 1 and 2 **Task 1** **Anchor task**	Council of Europe A2	Interview	Selecting multiple choice items
Booklet 1 **Task 2**	Basic	Dialogue	Matching places described with places on a map
Booklet 1 **Task 3**	Intermediate	TV interview	True/false questions
Booklet 2 **Task 2**	Intermediate	Part of a radio programme	Giving short answers (parts of the answers are given)
Booklet 2 **Task 3**	Advanced	TV interview	Giving shorts answers

Anchor task

Input text type: interview
Task type: selecting multiple-choice items
Task description and requirements: The candidates heard an interview with someone who works for The Tower of London. They were required to listen for and understand specific information in the text. Their task was to select multiple-choice items choosing from 3 options. The text could be heard only once. In order to complete the task successfully the candidates had to read extensively, as both the questions and the options were rather long. The text was played in parts and the candidates had to answer the questions in the pauses between the parts. The task was made more somewhat difficult by the fact that there was no time to read the questions and options in advance, but this was how the task had been administered in the CITO Project, and so it was important not to change anything, otherwise the comparison of results with that project would have been invalidated.

Table 12.2: Descriptive statistics for Anchor task, Booklets 1 and 2 (10 items)

	Booklet 1	Booklet 2
Number of students	244	269
Mean	5.086	5.123
Standard deviation	2.403	2.564
Reliability	0.643	0.699
Mean % Correct	51	51
Mean Item-Tot Corr	0.487	0.519

Table 12.3: Item-level results for Anchor task, Booklets 1 and 2

Item	1	2	3	4	5	6	7	8	9	10
Booklet 1										
Facility	48%	52%	53%	48%	40%	44%	50%	48%	72%	55%
Discrim	.48	.53	.50	.56	.48	.55	.18	.13	.42	.55
Logit	-.76	-1.17	-1.13	-.91	-.62	-.63	-.88	-.81	-2.05	-1.24
Booklet 2										
Facility	46%	56%	53%	50%	46%	42%	48%	46%	71%	56%

Discrim	.64	.58	.36	.69	.55	.53	.45	.20	.44	.55
Logit	-.76	-1.17	-1.13	-.91	-.62	-.63	-.88	-.81	-2.05	-1.24

Tables 12.2 and 12.3 show clearly the value of having anchor items when two different populations take two different booklets. The item facility values and discriminations vary across the populations, as do raw means, standard deviations and reliabilities, but the logit values are common, enabling the calibration of the experimental tasks.

Booklet 1
Task 2

Intended level: Basic
Input text type: dialogue
Task type: matching places described with places on a map
Task description and requirements: The candidates heard a dialogue between two people about what one of them was doing after work. They were required to listen for and understand specific information in the text. Their task was to match the five places mentioned in the dialogue with the places on the map. The text was heard twice. Before the first listening the candidates had 30 seconds to study the task. After the second listening they were given another 30 seconds to finalize their answers. The task required careful reading as well as there were several places of the same kind (restaurant, cinema, car park, etc.) on the map.

Table 12.4: Descriptive statistics for Task 2, Booklet 1

Number of items	5
Mean	1.660
Standard deviation	1.427
Reliability	0.618
Mean % Correct	33%
Mean Item Tot Corr	0.610

Table 12.5: Item-level results for Task 2, Booklet 1

Item	1	2	3	4	5
Facility	12%	35%	28%	45%	46%
Discrimination	.09	.66	.33	.67	.70
Logit value	1.92	-.20	.19	-.78	-.88

Booklet 1
Task 3

Intended level: Intermediate
Input text type: TV interview
Task type: True/false questions
Task description and requirements: The candidates heard an interview about someone's life. They were required to listen for and understand specific information. Their task was to decide if the 10 statements given were true or false. The text was heard twice. Before the first listening the candidates were given 30 seconds to familiarize themselves with the task. After the second listening they had one minute to complete the task. The text contained information that the candidates may not have been familiar with, which added to the difficulty of the task.

Table 12.6: Descriptive statistics for Task 3, Booklet 1

Number of items	10
Mean	6.082
Standard deviation	1.936
Reliability	0.481
Mean % Correct	61%
Mean Item Tot Corr	0.427

Table 12.7 Item-level results for Task 3, Booklet 1

Item	1	2	3	4	5	6	7	8	9	10
Facility	40%	85%	58%	62%	78%	37%	53%	65%	48%	82%
Discrimination	.26	.28	.16	.44	.30	.24	.41	.00	.39	.43
Logit value	-.52	-3.25	-1.35	-1.58	-2.53	-.33	-1.13	-1.73	-.86	-2.77

Booklet 2
Task 2

Intended level: Intermediate
Input text type: part of a radio programme
Task type: giving short answers
Task description and requirements: The candidate heard a part of a radio programme where some people were talking about a particular object. They were required to listen for and understand specific infomation. Their task was to give short answers to 9 questions. The first parts (sometimes only an article or a preposition) and in two items the endings of the answers were given. The text was heard twice. Before the first listening the candidates were given 30 seconds to study the questions. After the second listening they had 30 seconds to finalize their answers. The task required understanding and writing down one or two words as answers to the questions.

Table 12.8: Descriptive statistics for Task 2, Booklet 2

Number of items	9
Mean	1.862
Standard deviation	1.913
Reliability	0.704
Mean % Correct	21%
Mean Item Tot Corr	0.529

Table 12.9: Item-level results for Task 2, Booklet 2

Item	1	2	3	4	5	6	7	8	9

Facility	33%	7%	19%	23%	21%	3%	16%	26%	39%
Discrimination	.64	.18	.45	.42	.62	.07	.21	.53	.59
Logit value	-.03	2.44	.97	.64	.82	4.21	1.49	.45	-.36

Booklet 2
Task 3

Intended level: Advanced
Input text type: TV interview (the same recording as in Booklet 1, Task 2)
Task type: giving short answers
Task description and requirements: The candidates heard an interview about someone's life. They were required to listen for and understand specific information. Their task was to give short answers to eight questions. The text was heard once only. Before listening to the text the candidates were given 30 seconds to study the question. After listening they had one minute to finalize their answers. Although the text was the same as in Task 2 of Booklet 1, the task was completely different and so was the focus of the questions. In this task only four questions focused on relatively easily recognisable facts, the other four required some kind of interpreting of what was heard. The task was made more difficult by the fact that it contained information that the candidates may not have been familiar with.

Table 12.10: Descriptive statistics for Task 3, Booklet 2

Number of items	8
Mean	1.301
Standard deviation	1.274
Reliability	0.519
Mean % Correct	16%
Mean Item Tot Corr	0.478

Table 12.11: Item-level results for Task 3, Booklet 2

Item	1	2	3	4	5	6	7	8
Facility	4%	50%	21%	5%	3%	15%	2%	30%
Discrimination	.09	.38	.50	.09	.09	.35	.05	.43
Logit value	3.69	-.94	.79	3.14	3.56	1.28	3.60	.25

The statistical analysis of the tasks reveals some interesting tendencies. It is particularly useful to compare the performance of the two Listening Booklets since they included the same anchor task, ensuring that, with the help of IRT, the perfomance of different candidates on the different booklets, and thus the performance of the tasks themselves would be comparable. However, even the descriptive statistics show that these two booklets performed very differently.

Booklet 1, in the light of mean facility values, appears to have been easier (51%) than Booklet 3 (31%). It is important to note here, however, that the two booklets were taken by different candidates, hence mean facility values—being sample dependent—do not necessarily reveal the true picture. IRT-based statistics make direct comparison across booklets possible. Nevertheless, IRT-based difficulty measures support the facility based conclusion. The average difficulty estimate of items in the anchor task was –1.02 logits in both booklets, of course, but the average logit scores for the other tasks in each booklet clearly revealed that Booklet 2 was more difficult. Task 2 in Booklet 1 had an average difficulty of +0.05 logits, and Task 3 in the same booklet –1.605 logits, while in Booklet 2 Task 2 showed an average difficulty value of +1.181 logits, and Task 3 +1.921 logits.

We should hasten to add that the difference between the two booklets in terms of difficulty had been anticipated, indeed intended, as Booklet 1 included—apart from the Anchor task—a task for Basic and another one for Intermediate level, while Booklet 2 contained an intended Intermediate and Advanced level task beside the Anchor task. Surprisingly, however, Task 2 in Booklet 1, an intended Basic level task, was more difficult than Task 3, intended for Intermediate level. Also, if we compare Task 2 in Booklet 2, an intended Intermediate level task, to any of the tasks in Booklet 1, there appears to be a fairly large difference in logit scores.

In the light of these findings it seems logical to assume that level setting will require several more pilot versions in order to determine what exactly makes a task difficult. This is particularly important to note in relation to Task 3 in both booklets. Here the same text was used with different tasks, and, as the difficulty estimates show, one of them had the lowest, while the other one the highest average difficulty. This indicates clearly that overall task difficulty may to a large extent be determined by the task type and not by the text itself. Further piloting, however, is necessary to verify this assumption.

The second issue we can explore in the light of the statistics is how the two booklets performed in terms of their discriminative power. The average item discrimination figures indicated no particular problems here. Booklet 1 showed an average discrimination of .38, while Booklet 2 .42. Both of these figures indicate acceptable levels of item discrimination. Interestingly, Booklet 2, which proved to be more difficult, appeared to have been slightly better in terms of discrimination.

A third important criterion to examine is reliability. While Booklet 2 had an acceptable .82 as reliability coefficient, Booklet 1 showed a somewhat low .76. While lack of reliability can be attributed to various sources, the case of Booklet 1 offers a likely explanation.

Since the Anchor task had been piloted and found suitable earlier, it can most probably be excluded as a source of the problem here. Besides, it was common to both booklets, and it did not cause a problem in Booklet 2. Task 2 in Booklet 1, however, was made up of a mere five items, which may well have resulted in a reliability problem. Even more likely, however, is the assumption that the bulk of the problem was caused by Task 3, a true/false task. Such tasks tend to be problematic in general in terms of reliability. It seems that in this particular case, where the task performed unexpectedly in terms of difficulty anyway, reliability was seriously affected by the use of this test method. A major lesson to learn from this is that no matter how frequently this task type is used by teachers, it is unsuitable for measurement purposes.

Having compared the two booklets in terms of difficulty, discrimination and reliability, it is useful to compare the performances of Year 10 and Year 12 students on the two booklets.

Table 12.12: Listening booklets: facility values for tasks, Years 10 and 12 compared

Item number	Year 10 FV	Year 12 FV
Booklet One		
Anchor 1	39%	52%
Anchor 2	38%	59%
Anchor 3	46%	58%
Anchor 4	35%	53%
Anchor 5	31%	45%
Anchor 6	28%	50%
Anchor 7	44%	54%
Anchor 8	53%	45%
Anchor 9	64%	76%
Anchor 10	51%	57%
Task 2.1	6%	15%
Task 2.2	22%	41%
Task 2.3	25%	29%
Task 2.4	33%	49%
Task 2.5	38%	51%

Task 3.1	35%	43%
Task 3.2	83%	86%
Task 3.3	53%	60%
Task 3.4	64%	63%
Task 3.5	71%	82%
Task 3.6	39%	35%
Task 3.7	43%	58%
Task 3.8	67%	63%
Task 3.9	40%	52%
Task 3.10	79%	83%
Booklet Two		
Anchor 1	42%	46%
Anchor 2	62%	53%
Anchor 3	38%	60%
Anchor 4	39%	53%
Anchor 5	45%	46%
Anchor 6	41%	42%
Anchor 7	45%	49%
Anchor 8	43%	48%
Anchor 9	73%	71%
Anchor 10	58%	56%
Task 2.1	36%	32%
Task 2.2	9%	5%
Task 2.3	19%	18%
Task 2.4	19%	24%
Task 2.5	19%	21%
Task 2.6	0	4%
Task 2.7	23%	12%
Task 2.8	23%	27%
Task 2.9	45%	37%
Task 3.1	7%	3%
Task 3.2	51%	49%
Task 3.3	15%	23%
Task 3.4	3%	5%
Task 3.5	3%	3%
Task 3.6	8%	17%
Task 3.7	0	3%
Task 3.8	32%	28%

As expected, the majority of items were answered correctly by a larger proportion of Year 12 than Year 10 students. In many cases, however, the difference was minimal, and in the case of some items a reverse tendency could be observed. Such items could be found in all tasks, but it would be a mistake to conclude from this that these items are all problematic. It is particularly important to note that the Anchor task performed quite differently in this sense in the two booklets. In Booklet 1 anchor item 8 was answered correctly by a larger proportion of Year 10 students than by Year 12 candidates, while in Booklet 2 items 2 and 9 showed a similar problem. A few points need to be clarified, however.

First, the actual difference in proportions in such cases tended to be quite low, indicating that the seemingly reversed pattern may not be an actual one.

Second, the size of the samples of the two year groups was relatively low. Thus proportional differences may be explained by the unexpected behavior of a relatively low number of candidates. Considering the fact that many of the items were objectively scorable, guessing cannot be ruled out either. Hence, potentially, only a few Year 10 students' successful guessing may have produced the unexpected item facility values.

Third, as discussed in Chapter 8, the sample was not representative of Year 10 or Year 12 populations, and probably many Year 12 students who had passed a state language exam were not present.

Nevertheless, the fact that several items do not seem to make a difference between the two groups is worth noting. If true, this appears to imply that either the populations are not very different in terms of proficiency, or that several items need revision. The composition of the sample may explain the first point, and the second point may only be true for high difficulty tasks, as a very easy task may cause no particular problem for a large proportion of Year 10 students either.

In sum, the statistical results offer a chance to evaluate how the piloted tasks worked, but the conclusions we can draw are limited. Yet the data serve as invaluable assistance to future test development in all fields examined.

After an overview of the piloting along with a discussion of some of the most important findings, we now need to discuss what problems need solutions.

Issues to be resolved

A crucial issue is the delivery of the Listening Paper itself. While the actual booklets pose no particular problem, listening as a skill tested is special in the sense that sound quality must be guaranteed not only at the stage of task construction, but also at the stage of test delivery. It is of paramount importance to guarantee that candidates in various parts of the country have equal chances in this respect. Considering the fact, however, that, according to the findings of the Baseline Study, in several schools technical facilities for Listening tests are either not available or are of poor quality (Fekete et al 1999:239), this problem needs an urgent solution.

One solution proposed is to deliver the text through the media, either TV or radio, or both. The present version of the Specifications already makes mention of this possible solution, but further investigations into whether and how such facilities can be utilized are necessary. Apart from the sound quality problem, this solution would also reduce the likelihood of potential breaches of security, as the recording would only be available centrally.

A further issue that needs to be resolved is the definition of levels in terms of piloted tasks. As the first piloting has shown, expectations are sometimes inaccurate. In some cases this does not necessarily pose a major problem, as some texts and task types are common across levels. Other tasks, however, may need major revision or would need to be dropped if found unsuitable for the level they were intended for. Alternatively, the Specifications may need to be revised in the light of the findings to allow more common texts and tasks across the two levels.