Chapter 14

Use of English

Martsa Éva and Nyirő Zsuzsanna

Why test use of English and why only at Advanced level?

As we have seen, both the Basic Year 10 and the Year 12 School-leaving Examinations assess candidates' proficiency in the areas of the four skills (reading, listening, writing and speaking). In addition, at the Advanced level linguistic competence is assessed both indirectly (in testing writing and speaking, as at Intermediate level) and directly in a Use of English Paper, which aims to assess whether the candidate possesses the lexical, grammatical, semantic and pragmatic knowledge that will enable him/her to communicate independently. At the other levels linguistic competence is assessed in integrated tasks for the four skills. (*Specifications for the School-leaving Examinations in English*, October 1998)

There is, however, necessarily an overlap between the testing of reading and testing use of English. As Alderson says, 'In (teaching) reading, it is common to make a distinction between a student's ability to understand the syntax of the text, its vocabulary and the relations between and among sentences, and a student's ability to understand the main idea, to read between the lines of the text to infer indirect meaning, to recognise text types, and to do things like recognise the author's purpose, the author's bias, the difference between fact and fiction, relevant and irrelevant and so on.' (Alderson, 2000)

And indeed, comparing the requirements in the Specifications, some (sub)skills are required both in the Reading and Use of English Papers. Inferring meaning from context, demonstrating understanding of how information is structured in a text, recognizing the cohesive devices and their functions within sentences and in text or, understanding the role of particular punctuation marks are equally important for a candidate to perform successfully on both papers. In addition, some task types are similar (multiple-choice, multiple matching, banked or modified gap-filling). But while the emphasis in the Reading Paper is mainly on ***recognition*** of these structures / devices and on the ability to use them to understand the text, candidates are expected either to ***produce*** them in UE Papers, or the UE test is confined to testing candidates' ability to recognise *correct* structures, lexis, etc, but not to use them to construct overall meaning of text.

In order to distinguish as clearly as possible between testing reading and use of English, the revised Detailed Requirements are as follows:

For testing reading they say that

'the candidate should be able to

- *understand the gist of the text*
- *follow the chain of ideas in a text*
- *find specific information by scanning*
- *separate relevant and irrelevant information*
- *infer meaning from context*

- *demonstrate understanding of how information is structured in a text*
- *recognise the role of linguistic (lexical, grammatical) devices in creating text coherence.'*

For testing use of English the Detailed Requirements say that

'the purpose of the 'Use of English'Paper is to assess whether the candidate possesses the lexical, grammatical, semantic and pragmatic knowledge that will enable him/ her to communicate independently. Contrary to the specifications for the school-leavig exam in other foreign languages, this paper constitutes a part of the English exam at Advanced level only.'

The particular requirements related to this part of the exam include the ability to

- *recognise, complete, produce common and less frequently used syntactic structures*
- *use common and more sophisticated lexical items, taking into consideration syntactic features, collocations, prepositions*
- *use grammatical structures and lexical units in different text types.*

A part of the tasks may be similar to those of the Reading Paper, but the difference is that in this part of the examination (UE) the task focuses on the knowledge of grammatical and lexical elements and not on the global meaning expressed by the entirety of all these elements. Therefore the texts in this part of the examination are less complex so that problems with reading comprehension do not impede correct solution of the task.'

Introducing a separate paper for testing Use of English may be acceptable for several reasons. Even some large-scale proficiency tests (e.g. UCLES tests) retain a UE section to meet stakeholders' expectations. In addition, it can also serve some diagnostic purposes.

Direct testing of language skills instead of the linguistic abilities underlying them has resulted in the absence of any grammar component in proficiency tests. However, it can be argued that grammatical ability, or rather the lack of it, sets limits on what can be achieved in performing on skills-based tasks. So information on grammatical and lexical ability could be diagnostic and could be gathered through tests of grammar and vocabulary, not necessarily as an integral part of a reading test.

At lower levels, communication can still be successful even with poor grammatical ability. However, the higher the level a student is at, the more grammar and vocabulary can contribute to master communicative skills. Hence knowledge of the candidates' grammatical ability could be useful information.

One might argue that there is no need to justify the separate testing of use of English in Hungary. Grammar and translation have always played an important role both in teaching and testing foreign languages especially in the Matura (érettségi) examinations. With the introduction of modern textbooks, however, this practice in teaching is changing, although the process is slow and classroom procedures (and the attitudes of teachers) have not changed so much yet.

However, translation is not present in plans for the new School-leaving examination. The omission of translation can be justified on at least two grounds. First, translation as a task in an examination is highly controversial and there are more arguments against than for it. Secondly, marking translation is highly subjective and therefore produces poor reliability (see Fekete et al, 1999 pp. 26-27.) In addition, a positive washback effect can be expected – as there is no need for practising translations for the examination, teaching practices and classroom procedures are likely to change as well. It could be said that the same argument applies to the omission of tests of Use of English at Basic and Intermediate levels.

Apart from their diagnostic potential as discussed above, and in addition to the need not to introduce too many radical changes at once, a further reason for retaining tests of use of English is the potential contribution to objectivity and reliability. A large number of items can be scored objectively in a use of English test, and the fact that there can be so many items and there is no need for judgement when scoring may have some advantages.

Comparing previous practice in testing use of English with the planned reform

The present school-leaving examination in English is a large-scale semi-achievement test consisting of two parts. Though there are several versions of it, there are four basic types (all of which containing UE and/or translation tasks).

Table 14.1: Types of current school-leaving examination and the place of UE

matura	***Basic***	***Special***	***Bilingual***	***Joint-entrance***
written A	UE test*	UE test*	2 translation tasks (E>H, H>E)	UE test*, reading comprehension*
B	translation(E>H)	guided composition*	essay writing*	guided composition*
oral A	summarising unfamiliar texts (in Hungarian),*	summarising unfamiliar texts (in English),*	analysing unfamiliar texts (in English),*	text analysis (in Eng.),* pictures description, prepared topics, civilization, etc.
B	talk on prepared topics	talk on prepared topics	talk on prepared topics	

* Dictionaries are allowed.

The different parts of the school-leaving examination(s) focus on testing different things.

Table 14.2: Use of English tasks in the present matura

matura	***what it tests***	***scoring***
written part A	an indirect testing of vocabulary, grammar and functions and also some writing and limited reading comprehension	mainly objective
B	either a direct testing of writing (e.g. guided compositions or essays) or a highly integrative translation task (mediation) including testing grammar and functions, vocabulary, spelling, punctuation, register in two languages, and testing reading skills, problem solving, communicative and sociolinguistic competences, knowledge of the world, imagination as well (Noijons & Nagy 1996)	subjective
oral part A	a direct testing of reading comprehension and an indirect testing of some linguistic competence and mediation	subjective
B	a direct testing of oral proficiency (lacking authenticity and information gaps) and a very limited testing of listening comprehension in interaction	subjective

With the exception of the joint school-leaving- entrance examination, testing use of English is heavily structural in content and there are few examples of testing grammar meaningfully within a context. The majority of the tasks in the basic and special school-leaving examinations are sentence-based, testing usage and competence, not use and performance. They are examples of discrete-point testing, focussing on the so-called 'difficult aspects of grammar' (Noijons & Nagy, 1996).

The most typical task types are: multiple choice, sentence completion, gap-filling, paraphrasing, transformation (the beginning of the sentence is given), making questions / negative sentences, short answers, word derivation, replacing vocabulary, joining sentences, giving appropriate utterances (communicative), etc.

Multiple matching is rare, options provided in gap-filling tasks are always as many as there are gaps. There are no 'spot the error tasks' (identifying, correcting, deleting or supplying an omitted word). However, some of these tasks can be found in the joint-entrance examination tests.

It must be admitted that the school-leaving examinations (especially the basic and bilingual ones) do not meet the requirements mentioned either in the previous (1978) or the latest (1995) national curricula. For instance,

a) the present Matura does contain translation, but that skill is not part of the curriculum and students are not trained for this activity.
b) In the latest curriculum, (which is applied only to the first ten years, as additional regulations for the last two years are being worked on at the moment), grammar is treated as a subskill. However, it still dominates the examinations.
c) On the other hand, teaching the four language skills is required, but they are not equally tested and weighted in the examinations.

Testing Use of English in the new School-leaving examination system

As mentioned in previous chapters, the new School-leaving examination system in English will retain a test of the Use of English as a separate component at Advanced level, but it will not dominate. Linguistic competence will be tested both in sentence-based and text-based tasks. Items will sample widely from the structures specified, and should reflect a functional approach.

Sentence-based task types include multiple choice sentences (choosing the right answer from 4 options) and sentence transformation (where a part of the new sentence or the structure / word to be used is given).

Text-based tasks are as follows: multiple choice; inserting the appropriate form of a word into the text using a given stem; gap filling; banked gap filling; identifying errors; identifying and correcting errors (deleting words, supplying an omitted word, correcting incorrect lexical or grammatical elements); arranging jumbled sentences.

Text-based tasks and where possible sentence-based tasks as well are drawn from authentic texts.

Description of piloted test papers / UE tasks (April 1999)

In April 1999 four test booklets were piloted, following the piloting of several Speaking and Writing tasks in 1998. As described in Chapter 4, Booklets 1 and 2 consisted of Listening, Reading and Use of English tasks. Booklets 3 and 4 contained only Reading and Use of English items.

Table 14.3: Description of the Use of English tasks (summary)

booklet	***task title***	***items, length in words***	***task type***	***method of testing***	***focus on***
1.4	*Move Over, Webster**	15 228	spot the error (identifying and deleting)	text containing errors, extra words to be identified and deleted	proof-reading, applying knowledge of the language system, especially grammar
1.5	*A Short Story**	5 70	sequencing	short narrative text consisting of five jumbled sentences	cohesion, demonstrating understanding how text structure operates
1.6	*Migratory Birds*	33 333	spot the error (identifying)	long text containing errors, extra words to be identified	proof-reading, applying knowledge of the lang. System, especially grammar and lexis
2.6	*What on Earth...?*	15 209	gap-filling	text containing 15 gaps	testing lexical accuracy (prepositions), with more than 1 acceptable words for some gaps
3.8	*Spice Girls*	16 225	gap-filling	text containing 16 gaps	testing lexis and structure and cohesive devices
4.8	*(discrete sentences)*	10 166	spot the error (identifying + correcting)	identifying the incorrect elements in discrete sentences containing four underlined words /phrases	testing lexis and structure
4 .9	*(discrete sentences)*	11 213	multiple choice	text containing 11 gaps in discrete sentences followed by 11 four-option multiple choice answers	testing lexis and structure
4.10	*(discrete sentences)*	9 115	transformation	discrete items: prompt sentence and response sentence of which the beginning is given	testing lexis and structure

* Used as anchors for calibrating items

The first two Use of English tasks (*) were used as anchors for calibrating items across the various test booklets, but depending on what preceded them they were numbered differently in the different test booklets. In Reading Booklets 1 and 2 they were numbered 4 and 5 and in Reading Booklets 3 and 4 they were numbered 6 and 7. The first three booklets contained 3 UE tasks, Booklet 4 contained 5 tasks.

All the four booklets were taken by different groups of Year 10 and Year 12 students.

General comments

Task design, items

Although there were altogether eight Use of English tasks included in the piloted booklets, one of the anchors (a sequencing task) might be regarded as a Reading task since it requires textual comprehension as well as linguistic abilities.

Five of the tasks were text-based integrated tasks, three of them were sentence-based.

There were three production grammar tests (2 gap-filling tasks and a sentence transformation task), requiring the students to supply lexical and grammatical structures appropriately.

The proportion of lexical and structural items varied from task to task.

Layout was more or less standardised except that a) lines were not even in two tasks *(What on Earth, Spice Girls),* b) there was more than one item (extra word) in one line *(Migratory Birds),* c) the boxes for items in the latter task were not aligned with the lines and they were not numbered, nor was the maximum score indicated; d) in some cases items were at the end of the lines, contrary to the Guidelines for Item Writers.

There was only one task *(What on Earth)* which used drawings as illustrations, but as the order of them was the opposite of that mentioned in the text, this may have been confusing.

Text selection

Most of the texts were interesting and the topics were involving. Text-based tasks appear to have been constructed on authentic newspaper articles, although the sources of the texts were not always indicated *(A Short Story, Migratory Birds, Spice Girls).* The intended level of difficulty was, obviously, Advanced. However, the anchor sequencing task was well below Advanced level.

Except for one task *(Migratory Birds)* the length of texts did not exceed what is specified in the Guidelines for Item-writer (cc.300 words).

Rubrics

Rubrics were clear with the exception of the first discrete sentences UE task in Booklet 4. They were relatively brief and standardised.

Detailed description of the piloted use of English tasks and discussion of the empirical results

For samples of the booklets and tasks see Appendix IV.

Booklets 1, 2, 3, 4 – Anchor Task 1 (*Move over, Webster)*

Input text type: newspaper article
Number of items: 15
Focus on: structure (6), lexis (6), blank (3)
Length of rubrics and text: 60 / 228
Task type: spot the error and identify, text-based
Requirements: proof-reading

Task description: This task served as an anchor task in all the four Reading booklets, in order to be able to calibrate the remaining UE tasks, which only appeared in one booklet each. Candidates were expected to apply their knowledge of the language system in order to identify and delete extra words from an authentic text containing errors. The text was a 17-line long part of a newspaper article about Asian English. In 15 lines the candidates had to find and identify a possible unnecessary word. Some lines were correct, but the number of these was not given.

The layout is standardised and easy to handle (though items 4 and 6 are at the end of the lines). The task focusses on spotting errors both in grammar structures and in lexical elements.

The rubric is as brief as possible. There are 2 examples provided in order to make it clear that some of the lines are correct. The ratio between the number of items and the words in text (15 items for 228 words) is acceptable. The text does not challenge reading skills too much, but it is appropriately difficult for this level.

The reliability of the task is moderately high, ranging from 0.660 to 0.702 in the four booklets. It discriminated reasonably well (0.386-0.446).

Table 14.4: UE Anchor 1 – Task level analysis (Move over, Webster)

	Booklet 1	*Booklet 2*	*Booklet 3*	*Booklet 4*
Mean	3.651	3.593	2.912	2.090
Std.dev.	2.620	2.611	2.010	1.940
Alpha	0.714	0.702	0.589	0.660
Mean Pcnt Corr	24 %	24 %	19 %	14 %
Mean Item-Tot.	0.446	0.445	0.386	0.406

The IRT data prove that this task was a fairly difficult one, the mean logit value being +1.891.

Students reported taking an average of 10 minutes to complete this task, ranging from a minimum of a few seconds to a maximum of 50 minutes.

A detailed description of what each item tests:

Item 1	unsuitable use of an object
Item 2*	unsuitable lexical item
Item 3	unsuitable preposition
Item 4	unsuitable use of an auxiliary
Item 5	misuse of an adverb
Item 6	wrong tense
Item 7	correct line
Item 8	misuse of passive
Item 9	unsuitable lexical item
Item 10	correct line
Item 11	unsuitable lexical item
Item 12	misuse of an article
Item 13	unsuitable lexical item
Item 14	misuse of passive
Item 15	correct line

(* Item 2 originally was meant to be a correct line but during marking it turned out that deleting a word was acceptable.)

Table 14.5: Anchor Task 1 – Facility value and discrimination index (Move Over, Webster)

No. of item	*FV B 1 in %*	*FV B 2 in %*	*FV B 3 in %*	*FV B 4 in %*	*DI B 1*	*DI B 2*	*DI B3*	*DI B4*	*Logit value*	*Item focus*
1	**12**	**9**	**6**	**4**	.32	***.13***	***.21***	***.08***	2.94	str
2	53	45	42	33	.50	***.22***	.39	***.26***	-.01	lexis
3	31	30	18	14	.68	.40	.46	.32	1.28	lexis
4	**13**	**13**	**6**	**4**	***.27***	***.16***	.38	***.07***	2.73	str
5	**17**	**19**	**9**	**4**	***.27***	***.22***	.41	***.10***	2.14	str
6	**8**	**8**	**2**	**1**	***.07***	***.06***	***.05***	***.03***	3.91	str
7	63	63	53	44	.53	.47	.44	.34	-.61	-
8	**19**	**20**	**8**	**7**	.43	***.24***	.41	***.21***	2.22	str
9	**9**	**11**	**3**	**3**	***.12***	***.09***	***.26***	***.06***	3.35	lexis
10	37	36	33	34	***.15***	***-0.3***	***.17***	***.18***	.46	-
11	**10**	**11**	**3**	**3**	***.23***	***.09***	***.26***	***.10***	2.96	lexis
12	23	28	**9**	**8**	.44	***.28***	.53	***.21***	1.73	lexis
13	**12**	**16**	**8**	**3**	.35	***.26***	.39	***.08***	2.42	lexis
14	**8**	**9**	**3**	**4**	***.25***	***.04***	***.23***	***.10***	3.17	str
15	51	43	54	43	***.20***	***.04***	***.07***	.33	-.22	-

FV – facility value DI – discrimination index B 1, 2, 3, 4 – Booklets 1, 2, 3, 4

The data for the items that proved to be too difficult are printed in bold while the data for the items that did not discriminate well are printed in bold italics.

In all four groups candidates performed best on Items 2, 7, 10 and 15. Lines 7, 10 and 15 were correct ones, in line 2 there was an unnecessary lexical item (as mentioned above this one was also meant to be a correct line). Finding the correct lines proved to be the easiest subtask

Items 1, 4, 5, 6, 8, 9, 11,13 and 14 proved to be too difficult. In Items 9, 11 and 13 candidates were required to find an unsuitable lexical item. This subtask proved to be the most difficult, three out of the four items having very low facility values.

Regarding discrimination, nearly half of the items (1, 4, 6, 9, 10, 11, 14 and 15) did not discriminate well.

Booklets 1, 2, 3, 4 – Anchor Task 2 *(A short story)*

Input text type: a short anecdote
Number of items: 4/5
Focus on: cohesion
Length of rubrics and text: 49 / 70
Task type: sequencing (arranging jumbled sentences)
Requirements: demonstrating understanding text cohesion

Task description: This task also served as an anchor. There were two versions slightly differing from each other. The text consisted of five sentences. The sentences were jumbled and candidates had to put them into the right order. In Booklets 1 and 2 the first sentence was given, in Booklets 3 and 4 all the five sentences were mixed up. That is why there were altogether five items in Booklets 3 and 4, but only four items in Booklets 1 and 2.

Some would argue that this is more of a Reading task than a Use of English one. The items focussed on cohesive devices and candidates were required to recognise the grammatical structures that made the text coherent.

Table 14.6: UE Anchor 2 – Task level analysis (A short story)

	Booklet 1	*Booklet 2*	*Booklet 3*	*Booklet 4*
Mean	2.488	2.834	2.298	2.145
Std.dev.	1.634	1.497	1.870	1.875
Alpha	0.879	0.860	0.816	0.818
Mean Pcnt Corr	62 %	71 %	46 %	43 %
Mean Item-Tot.	0.853	0.831	0.753	0.762

The level proved to be well below Advanced. The task was easy for this population, the mean logit value was -0.635 in Booklets 1 and 2, and – 0.668 in the other two booklets and the mean facility value ranged from 43 to 71 in the four booklets. But it was reliable (0.818 – 0.879) and discriminated well (0.762-0.853).

Students reported taking an average of 4 minutes to complete this task, ranging from a minimum of a few seconds to a maximum of 25 minutes.

Table 14.7: Anchor Task 2 – Facility value and discrimination index (A Short Story)

No. of item	*FV B 1 in %*	*FV B 2 in %*	*FV B 3 in %*	*FV B 4 in %*	*DI B 1*	*DI B 2*	*DI B3*	*DI B4*	*Logit value*
1	55	65	40	38	.68	.49	.59	.54	-0.40
2	-	-	52	51	-	-	.51	.61	-0.80
3	57	65	44	40	.73	.54	.68	.62	-0.49
4	78	86	43	42	.55	.44	.57	.66	-1.00
5	59	67	50	43	.67	.47	.60	.57	-0.65

Booklet 1 – UE Task 3 *(Migratory birds)*

Input text type: magazine article
Number of items: 33
Focus on: structure (21), lexis (12)
Length of rubrics and text: 54 / 333
Task type: spot the error (identifying), text-based
Requirements: proof-reading

Task description: Candidates were supposed to apply their knowledge of the language system in order to identify and delete extra words from an authentic text containing errors.

The layout is not standardised and was difficult for markers to handle. Lines were not even. The boxes for items are not numbered and they are not aligned with the lines. There are more than one item per line and there are too many items (33) for one task, even though the ratio between the number of items and words in text is 1 to 10. Two thirds of the items focus on grammar, 12 items focus on vocabulary. The range is not very wide, and a greater variety of foci would have been preferable.

The text is presumed to be authentic, but the source is not indicated. The level of difficulty of the text may be appropriate for this level, but the task challenges reading skills somewhat more than desirable. This is partly because of the length of the text (333 words altogether), contrary to the Guidelines for Item Writers, and partly because of the special vocabulary containing some technical terms concerning aviation (such as *'process genuine aircraft returns', 'transponders', 'altitude', 'blips', 'secondary radar'*, etc.). As there are a number of composite sentences (with several sub-clauses) and there are not enough words in between some items, it is difficult to restore the original text.

Detailed description of what each item tests:

Item	1	wrong aspect (perf. inf. instead of simple)
Item	2	unnecessary adverb
Item	3	unnecessary preposition
Item	4	wrong negative form
Item	5	unnecessary preposition
Item	6	countability
Item	7	unnecessary auxiliary
Item	8	unnecessary preposition
Item	9	wrong voice (passive instead of active)
Item	10	unnecessary object (pronoun)
Item	11	wrong voice (passive instead of active)
Item	12	unnecessary preposition
Item	13	wrong tense
Item	14	unnecessary preposition
Item	15	wrong negative instead of affirmative

Item	16	unnecessary preposition
Item	17	wrong conjunction
Item	18	unnecessary preposition
Item	19	unnecessary preposition
Item	20	wrong article (wrong number)
Item	21	unnecessary preposition
Item	22	unnecessary preposition
Item	23	unneces. article (specific instead of general)
Item	24	wrong auxiliary
Item	25	unnecessary object / pronoun
Item	26	unnecessary preposition
Item	27	wrong voice (passive instead of active)
Item	28	unnecessary article
Item	29	unnecessary object / pronoun
Item	30	wrong voice (passive instead of active)
Item	31	'to' infinitive instead of a plain infinitive
Item	32	unnecessary article
Item	33	unnecessary number

The task is highly reliable (.91), probably due to the number of items. The mean facility value is 35.9. Detailed item analysis is not possible, however, due to the problems marking this task (see Chapter 6).

Students reported taking an average of 12 minutes to complete this task, ranging from a minimum of 1 minute to a maximum of 24 minutes.

Booklet 2 – UE Task 3 *(What On Earth…?)*

Input text type: magazine article
Number of items: 15
Focus on: lexis
Length of rubrics and text: 30 / 209
Task type: gap filling (prepositions)
Requirements: using lexical units and cohesive devices

Task description: This task is a text-based production UE task, testing lexical accuracy, mainly prepositions. The text was an extract from an article about an unusual device. Visual prompts were given, showing the device described in the article – a pair of sugar nippers. The candidates had to fill in 15 gaps with suitable prepositions (with one suitable word for each).

While marking this test it turned out that 5 items (Items 2, 3, 9, 12, 14) had two acceptable solutions and Item 13 had even more (these were of course incorporated into the mark scheme). Most of the items focussed on vocabulary (prepositions, adverbs or particles in phrasal verbs, etc.).

The task was designed with a clear layout on facing pages. There is enough context provided in between the gaps and one line contains only one item, which is numbered. The example (0) ought to have been given in the text, not in the subtitle. Three items (Items 3, 6 and 12) are at the end of the line, which is contrary to the Guidelines for Item Writers. Examining the layout it becomes evident that the item writer deliberately ignored producing even lines, otherwise there might have been two items on one line (one of which would then have had to be dropped).

The text is authentic. The topic is slightly puzzling but it does not challenge reading skills very much. On the other hand, the vocabulary may have caused some difficulty. This might have been the reason why pictures / illustrations were used as well. However, they are somehow confusing, since they are not presented in the order in which they are mentioned in the text.

Table 14.8: Booklet 2, UE Task 3 – Task level analysis (What on Earth…?)

Mean	4.743
Std.dev.	3.639
Reliability	0.844
Mean Pcnt Corr	32%
Mean Item-Tot Corr	0.545

The task is fairly reliable (0.844), the discrimination index (0.545) is good. But the mean facility value is low (32%).

Students reported taking an average of 10 minutes to complete this task, ranging from a minimum of a few seconds to a maximum of 44 minutes.

Table 14.9 Booklet 2, UE Task 3 – Facility value and discrimination index (What on Earth…?)

No. of item	*FV in %*	*DI*	*Logit value*
1	**1**	***.03***	5.54
2	55	.62	-.15
3	**17**	.43	2.10
4	61	.66	-.53
5	26	.56	1.51
6	48	.69	.18
7	**16**	***.23***	2.36
8	33	.59	1.06
9	**18**	.36	2.00
10	**12**	***.26***	2.81
11	57	.77	-.22
12	35	.66	.91
13	36	.66	.86
14	34	.59	1.04
15	25	.48	1.59

There were five items that proved to be too difficult – printed in bold, out of which three did not discriminate well either – printed in bold italics. Item 1 required the preposition ***over*** with the meaning of 'beyond some quantity'; Item 3 needed a preposition of place but it was difficult to decide which one without being familiar with the device; Item 7 focussed on the use of ***in*** meaning 'during'; Item 9 and 10 were difficult for those who did not the know the words following the prepositions (*'slab'* and *'genteel lumps'*).

Out of these five items three did not discriminate well (1, 7, 10).

The IRT data prove that the task was moderately difficult, the mean logit being +1.404.

Booklet 3 – UE Task 3 *(Spice Girls)*

Input text type: newspaper article
Number of items: 16
Focus on: lexis (10), structure (2), cohesion (syntax) (4)
Length of rubrics and text: 26 / 225
Task type: gap filling (suitable words)
Requirements: using grammatical structures, lexical units, cohesive devices

Task description: This task is a production UE task, testing lexis, structure and cohesive devices. It is based on an authentic text containing 16 gaps. The source of the text is not indicated.

In this task the candidates were supposed to apply knowledge of the language system and produce an acceptable word for each gap. 16 words were deleted from a 21-line long text and candidates were required to fill in the gaps with suitable words.

There is some inconsistency with the Guidelines (e.g. the length of the lines is uneven and item 11 is at the end of the line).

There is a variety of items focussing on different things (2 items on structure, 10 items on vocabulary and 4 items on syntax). Two of the vocabulary items (10, 12) test conjunctions and so more than one third of the items focus on special cohesive devices.

Table 14.10: Booklet 3, UE Task 3 – Task level analysis (Spice Girls)

Mean	2.168
Std.dev.	2.792
Reliability	0.826
Mean Pcnt Corr	14%
Mean Item-Tot Corr	0.539

The task is reliable (0.826) and discriminated well (0.539) but it seemed too challenging for this student population (+ 1.87) and the facility value is very low (14%).

There might be several reasons for this:

a) certain elements of the vocabulary may have been unfamiliar (e.g. *'laden', 'token', 'goodie bags', 'hug', 'besotted'*, etc.), although there were not so many that they would hinder understanding

b) the relationship to the topic may be ambiguous (Spice Girls and Prince Charles)

c) there were several complex sentences (and hidden sub-clauses) so the text challenged reading skills

d) problems with designing the task and items, namely that there were too many cohesive devices deleted from the text in some places so the remaining text (having additional gaps testing vocabulary or structure) might have lacked enough cohesion to be restorable.

e) Students seemed to get more and more tired while working on this booklet and this UE task was at the end. Booklet 3 was not the longest one (cc.3260 words including rubrics), but it contained only Reading tasks (5, cc. 2600 words) and UE tasks (3, cc.660 words). The student population taking Booklets 1 and 2 seemed to be more capable according to the statistics. However, those booklets tested Listening, Reading and Use of English and having different skills tested in one booklet might be less monotonous and might result in higher scores.

Students reported taking an average of 13 minutes to complete this task, ranging from a minimum of a few seconds to a maximum of 50 minutes.

Detailed description of what each item tests and their statistical data:

Table 14.11: Booklet 3, UE Task 3 – Facility value and discrimination index (Spice Girls)

No. of item	*FV in %*	*DI*	*Logit value*
1	**6**	***.12***	2.87
2	**7**	***.24***	2.21
3	**0**	***.00***	6.13
4	**7**	***.22***	2.16
5	**18**	.42	1.19
6	25	.70	.68
7	**9**	***.26***	2.06
8	**1**	***.03***	2.74
9	**14**	.40	1.50
10	22	***.5***	.88
11	**14**	.32	1.54
12	23	.32	.85
13	**11**	.31	1.71
14	**9**	***.24***	2.13
15	23	.59	.76
16	28	.69	2.1

No. of item	*Item content*
Item 1	verb (in Simple Past Tense)
Item 2	preposition
Item 3	nominal clause
Item 4	gerund
Item 5	formal subject
Item 6	passive (Nominative +Infinitive)
Item 7	preposition
Item 8	noun
Item 9	preposition
Item 10	conjunction
Item 11	preposition (passive, 'agent')
Item 12	conjunction
Item 13	prepositional verb
Item 14	preposition
Item 15	relative pronoun
Item 16	preposition

As the facility value data show, most items – 11 out of 16 – proved to be very difficult for the candidates. Even if we take into consideration that this population may have been somewhat less competent than those of Booklets 1 and 2, this task is clearly well beyond the target level.

We cannot explain why otherwise 'easy' items (1, 7, 8, 14) did not work well. The task could have worked better if it had focussed mainly on cohesive devices (in carefully selected places). The most difficult item (Item 3) focussing on syntax (nominal clause or subject clause) had 0 % facility value and 6.13 logit value. Thus it could not discriminate at all. In addition, eight of the sixteen items seem not to discriminate well.

The IRT data also show that the task was very difficult, the mean logit is +1.87, nearly as high as that of Anchor Task 1.

Booklet 4 – UE Task 3 *(Discrete sentences 1)*

Input text type: discrete sentences, not linked with each other in any way
Number of items: 10
Focus on: lexis (7), structure (3)
Length of rubrics and text: 15 / 166
Task type: multiple-choice 'spot the error and correct'
Requirements: identifying and correcting the incorrect elements
Task description: The task consisted of 10 sentences, each containing four phrases underlined. The candidates had to find the incorrect ones and correct them.

Table 14.12: Booklet 4, UE Task 3 – Task level analysis (discrete sentences 1)

Mean	0.791
Std.dev.	1.388
Alpha	0.710
Mean Pcnt Corr	8%
Mean Item-Tot.	0.522
Mean logit value	+ 1.95

Students reported taking an average of 8 minutes to complete this task, ranging from a minimum of 1 minute to a maximum of 40 minutes.

Detailed description of what each item tests and their statistical data:

Table 14.13: Booklet 4, UE Task 3 – Facility value and discrimination index (discrete sentences 1)

No. of item	*FV B4 in %*	*DI B4*	*Logit value*
1	**9**	***.20***	1.89
2	**2**	***.07***	1.99
3	**3**	***.07***	3.27
4	**9**	***.25***	1.68
5	**8**	***.20***	1.94
6	**1**	***.03***	NA
7	**15**	.44	1.32
8	**6**	***.18***	2.09
9	**20**	.47	0.91
10	**6**	***.13***	2.46

No. of item	*Item content*
Item 1	noun instead of an adjective
Item 2	gerund instead of a noun
Item 3	wrong past participle
Item 4	wrong conjunction
Item 5	deixis (wrong adverb)
Item 6	adjective instead of a verb
Item 7	infinitive instead of a gerund
Item 8	adjective instead of a noun
Item 9	gerund instead of an infinitive
Item 10	wrong tense

All 10 items proved to be too difficult for the candidates. Although the population taking this test was less competent than those of Booklets 1 and 2, the level of the task is clearly beyond the intended level. The IRT data also show that this task was the most difficult of all the seven piloted tasks, the mean logit value being +1.95.

Also, only two out of 10 items discriminate well.

7 out of 10 items focussed on lexis (1, 3, 5, 6, 8) or a mixture of lexis and structure (2 and 4), and only 3 focussed on structure (7, 9 and 10).

Correcting wrong structures may be easier than recognising misused vocabulary. The logit values of vocabulary items ranged from 1.68 to 3.27, whereas structure items ranged from 0.91 to 2.46. Consequently, the two easiest items (7 and 9) were structure items. However, they discriminate well. Year 12 students performed better on these items (FV 17 and 23) than Year 10 students (FV 12 and 13).

Why Item 10, on the other hand, proved to be one of the most difficult items cannot be clearly explained. Interestingly enough, Year 10 students performed better than the Year 12 ones (FV 8 to 4).

It is also difficult to tell why Item 3 proved to be the most difficult in this task (3.27). Both Year 10 and 12 students must be familiar with the past participles of the most common verbs. Perhaps they are not aware which of the two verbs (*'lie'* or *'lay')* is transitive. Item 6 also has the lowest facility value (1) and the lowest discrimination index (.03) – *'enable'* may be a rare lexical item for this population.

It is worth having a closer look and trying to find the reasons for this task being the most difficult one of all the piloted UE tasks. One important reason stated above is that the population taking Booklets 3 and 4 was less competent than the population taking Booklets 1 and 2.

But there may have been some other reasons as well, for instance:

a) Secondary school students may not have been not familiar with this task type even though it is frequently used in university admission tests.

b) The language of the task is formal (see sentences 1, 2, 3, 4, 6, 8, 9) and may have challenged reading skills more than is desirable.

c) The editing of this task and the wording of the rubrics was not satisfactory. Whether it was deliberate or accidental on the part of the item writer is not known, but underlying certain elements that students were asked to concentrate on seems confusing and misleading. The rubrics say, *'One of the underlined phrases in each sentence is incorrect. Write down the structure correctly.'* In the example given we can find real 'phrases'. However, in the task itself several underlined distractors do not qualify as phrases in the grammatical sense because they do not constitute a syntactical unit. They are just randomly underlined collocations of words, for instance *'land or'* (item 4), *'now is'* (item 5), *'been living', 'poverty for'* (item 7), *'on Friday so'* (item 9). What is more, in items 4 and 5 the underlined structures to be corrected cannot be identified. *(both ... or, now ... there).*

d) As the task type was unfamiliar, students may have used an inapropriate strategy. They may have tried to find the wrong phrases without taking the whole sentence into consideration. Either because of this or because of the formatting of the text, they may have had no meaningful context to rely upon.

e) They may have tried to find wrong *structures* (as given in the rubrics) and did not expect misused *vocabulary* (although it was given in the example). It would be better if the number of items focussing on structure and lexis in UE tasks were balanced.

f) It is not only a 'spot the error' task as correction of the wrong elements is also included, and producing is always more difficult than recognising.

g) Last but not least it must be added that Booklet 4 seemed to be the most difficult booklet for the candidates to work on:

- it was the longest booklet (3755 words including rubrics and texts both in reading and UE tasks, 939 words in UE alone),
- it contained the most tasks (5 Reading tasks and 5 UE ones, 10 altogether),
- it consisted of difficult and unfamiliar task types (several multiple matching tasks, sentence transformation, 'identifying and correcting errors'), and it contained the largest number of difficult tasks (6 out of 10).

Booklet 4 – UE Task 4 *(Discrete sentences 2)*

Input text type: discrete sentences, not linked with each other in any way
Number of items: 11
Focus on: structure (8), lexis (3)
Length of rubrics and text: 11 / 213
Task type: multiple-choice
Requirements: recognising correct elements
Task description: The candidates were given 11 gapped sentences. They had to choose the right answers from four options.

Table 14.14: Booklet 4, UE Task 4 – Task level analysis (discrete sentences 2)

Mean	2.970
Std.dev.	1.678
Reliability	0.310
Mean Pcnt Corr	27%
Mean Item-Tot.	0.354
Mean logit value	+ 0.515

This UE task is not very reliable (0.310). This is probably because of the small number of items.

Students reported taking an average of 7 minutes to complete this task, ranging from a minimum of 1 minute to a maximum of 21 minutes.

Detailed description of what each item tests and their statistical data:

Table 14.15: Booklet 4, UE Task 4 – Facility value and discrimination index (discrete sentences 2)

No. of item	*FV B4 in %*	*DI B4*	*Logit value*
1	**9**	***.10***	2.29
2	43	***.16***	-0.53
3	**17**	***.13***	1.16
4	24	.38	0.66
5	38	.48	-0.24
6	23	.69	0.69
7	41	***.25***	-0.39
8	35	***.10***	-0.15
9	35	***.26***	-0.10
10	**18**	***.10***	0.91
11	**13**	***.10***	0.36

No. of item	*Item content*
Item 1	used to / be used to
Item 2	lexical item
Item 3	inversion (*not only ... but*)
Item 4	double reported question
Item 5	Future Perfect
Item 6	lexical item
Item 7	relative pronoun
Item 8	subjunctive
Item 9	lexical item and verb tense
Item 10	passive causative
Item 11	lexical item

Four items were difficult for the candidates (*used to, be used to*, inversion after '*not only*', passive causative and a lexical item concerning youth). All the others were relatively easy, as the IRT data show this task was the second easiest of the seven tasks with a mean logit value of +0.515. Eight items of the eleven, however, do not discriminate well. The most difficult one was Item 1 (2.29), possibly not because of the structures in focus (*used to / be used to)* but because the distractors discriminated too well.

Booklet 4 – UE Task 5 *(Discrete sentences 3)*

Input text type: discrete sentences not linked with each other in any way
Number of items: 9
Focus on: structure (syntax)
Length of rubrics and text: 12 / 115
Task type: sentence transformation
Requirements: applying knowledge of the language system (transforming)
Task description: The candidates were given 9 prompt sentences and they had to transform them. The beginnings of the responses were given.

Table 14.16: Booklet 4, UE Task 5 – Task level analysis (discrete sentences 3)

Mean	1.098
Std.dev.	1.716
Reliability	0.777
Mean Pcnt Corr	12%
Mean Item-Tot.	0.585
Mean logit value	+ 1.812

This was a reliable but fairly difficult task with a mean logit number of +1.812. The discrimination index is acceptable but the facility value was very low.

Students reported taking an average of 7 minutes to complete this task, ranging from a minimum of a few seconds to a maximum of 50 minutes.

Detailed description of what each item tests and their statistical data:

Table 14.17: Booklet 4,UE Task 5 – Facility value and discrimination index (discrete sentences 3)

No. of item	*FV in %*	*DI*	*Logit value*
1	27	.58	0.41
2	**2**	***.04***	3.80
3	**20**	.45	0.91
4	**3**	***.06***	2.98
5	**14**	.35	1.43
6	**12**	.32	1.60
7	**11**	***30***	1.57
8	**5**	***.15***	2.29
9	**15**	.35	1.32

No.of item	*Item content*
Item 1	unless / if
Item 2	Nominative + perfect infinitive
Item 3	reported question
Item 4	subjunctive
Item 5	optative sentences
Item 6	probability, inversion
Item 7	double reported question
Item 8	co-ordination / sub-ordination (and, while)
Item 9	probability

Seven items out of nine proved to be too difficult for this test-taking population. Three items (2,4,8) did not discriminate well, being those most difficult for the candidates. However the discrimination indices of three further items are not very high either. But even with the 'easier' items the facility values are low (27 and 20).

The reasons for the weak performance on this task may be various:

a) The student population taking Booklets 3 and 4 proved to be less capable according to the statistical data, as mentioned before.
b) Sentence transformation is a production task and producing is always more difficult than recognising.
c) All the items tested complex cohesive devices focussing mainly on clausal syntax (and not on phrasal syntax). Students were expected to transform structurally complicated sentences into syntactically more complex ones.
d) Booklet 4 was the longest and most difficult test booklet, including 5 Reading and 5 UE tasks, containing *all the three* discrete sentence subtests.

Discussion of data from statistical analysis (summary)

Table 14.18: Use of English tasks in all booklets. Summary of descriptive statistics

booklets tasks	*tasks*	*types*	*items*	*length rubr/text*	*reliability*	*mean pcnt corr*	*mean item tot.*	*Mean logit v.*
1.4	*Move Over, Webster*	spot the error (identifying)	15	60 + 228	0.714	24	0.446	+1.891
1.5	*A Short Story*	sequencing	5	49 + 70	0.879	62	0.854	-.0.635
1.6	*Migratory birds*	spot the error (identifying)	33	54 + 333	0.91	36	no data	no data
2.4	*Move Over, Webster*	spot the error (identifying)	15	60 + 228	0.702	24	0.445	+1.891
2.5	*A Short Story*	sequencing	5	49 + 70	0.860	71	0.831	-0.635
2.6	*What on Earth*	gap-filling (open cloze)	15	30 + 209	0.844	32	0.545	+1.404

3.6	*Move Over, Webster*	spot the error (identifying)	15	60 + 228	0.589	19	0.386	+1.891
3.7	*A Short Story*	sequencing	5	49 + 70	0.819	46	0.762	-0.668
3.8	*Spice Girls*	gap-filling (open cloze)	16	26 + 225	0.826	14	0.539	+1.870
4.6	*Move Over, Webster*	spot the error (identifying)	15	60 + 228	0.660	14	0.406	+1.891
4.7	*A Short Story*	sequencing	5	49 + 70	0.818	43	0.762	-0.668
4.8	*(discrete sents)*	spot the error (correcting)	10	15 + 166	0.710	8	0.522	+1.95
4.9	*(discrete sents)*	multiple choice	11	11 + 213	0.310	27	0.354	+0.515
4.10	*(discrete sents)*	Transformati on	9	12 + 115	0.777	12	0.585	+1.812

Final comments on the Use of English data

Comparing the results with those of Listening and Reading tasks, it appears that the piloted Use of English tasks are highly *reliable*. The reason for this might be found in the objectivity of scoring. Marking the test was carried out in team work, with the help of marking schemes (see Chapter 6). These were easy to use, even with 'spot the error' task types. It is more likely, however, that this is due to the number of items tested, since reliability is almost always a function of the number of items in a test.

The UE tasks had *the lowest mean percentage* of correct answers as they were among the most difficult tasks. The *discrimination* indices varied from booklet to booklet, but in general they show that these UE tasks discriminated quite well.

The only exception was one of the sub-tests *(A Short Story)*, even though there are doubts as to whether this is a UE task rather than a Reading task. It is highly reliable, its discrimination index was high. But it is easy for this population. The other anchor sub-test (*Move Over, Webster)* was more difficult. Its reliability was quite high, but the task did not discriminate as well as other tasks.

The results in Booklets 1 and 2 are better than in the other two. Remember that these were different populations. This means that the students taking Booklets 3 and 4 were generally of lower ability than those taking Booklets 1 and 2.

Results in testing Reading and UE are said to correlate. (See Chapter 8 for the data on this.) However, the reason why Reading was easier in this pilot than UE may have been because scanning for specific information, inferring meaning from context or recognising certain cohesive devices between sentences, paragraphs, etc. was easier for these students than producing certain structural elements. Students performed relatively better on the skills-based tests.

The *new text-based UE task types* were challenging as they may have been quite unfamiliar to these students. But they worked relatively better than the sentence-based items. The fact that the *discrete sentence-based items* were on the whole more difficult may suggest either that such test methods may give a false picture of candidates' abilities, or that the tasks were not well designed. In addition, tasks requiring production instead of recognition always prove more difficult.

Expectations of stakeholders concerning testing Use of English

The success of the examination reform will depend on a whole range of factors. Concerning Use of English we must not forget that some task types are new, despite the use of modern textbooks which use such exercise types. As test items, they may not yet be familiar to students and above all to teachers.

Given the strong tradition of discrete-point testing of structures at all levels in Hungary, it will be very important to explain why there is no testing UE at lower levels. On the other hand, UE test papers at Advanced level will need to be sufficiently difficult for universities to consider the Advanced Matura acceptable as part of their entrance requirements.

The reform requires new ways of teaching UE in classrooms. Instead of concentrating on form, function and meaning should be stressed. Teaching how to use cohesive devices is also essential.