Chapter 8

Results of Piloting

Charles Alderson, Szabó Gábor and Pércsich Richárd

In this chapter, we give the main results from the piloting of the Writing tests, and the Written examination. The (relatively sparse) statistical results of the piloting of the Speaking tests are given in Chapter 11. In addition, more detailed analyses of each component of the Writing and the April Written tests are given in the relevant chapter (Chapters 10 and 12 to 14).

Results of marking the Writing tasks

As detailed in Chapter 6, a total of 260 scripts were marked by two raters. The main results are as follows.

Reliability

Since each script was marked by two raters (not, of course, always the same people), it was possible to calculate inter-rater reliability. For the test as a whole, regardless of booklet and excluding Booklet 1 for technical reasons (thus, for Booklets 2, 3 and 4 taken together), the overall correlation between the first rater and the second rater was a fairly respectable .83 (n=171, because of missing data). If we compare interrater correlations for Task 1, regardless of booklet, with similar correlations for Task 2, the coefficients are, respectively, a rather low .77 for Task 1, and a more reassuring .84 for Task 2.

These correlations are somewhat misleading, since they are aggregated over different tasks in different booklets. If we look at the inter-rater reliability for the total scores given by first and second raters, task by task, the results are as follows (Table 8.1):

Table 8.1: Correlations between First and Second Rater, Total Scores, by Task

Booklet	Task 1	Task 2
1	NA	.88 n=65
2	.84 n=67	.86 n=52
3	.76 n=62	.83 n=59
4	.73 n=60	.75 n=62

It would thus appear that some tasks are more reliably marked than others.

However, what these statistics hide is the fact that the ‘first rater’ was not always the same person, nor was the second rater. Thus, to see how different individuals agreed with each other, we have to compare scores given by particular pairs of raters. Unfortunately, because of the way the booklets were distributed, the number of scripts rated by any one pair varied considerably, and thus the correlations vary in their meaningfulness.

Table 8.2: Inter-correlations of total scores given by pairs of raters (all correlations significant at p<.05)

Raters	Task 1	Task 2
5:10	.93	.86
	n=18	22
3:5	.90	.97
	n=18	n=23
1:14	.92	.78
	n=10	n=11
2:7	.75	.85
	n=36	n=48
6:9	.86	.88
	n=34	n=45
4:14	.89	.88
	n=5	n=11
8:11	.81	.89
	n=38	n=39
3:10	.88	.74
	n=15	n=18
1:4	.67	.85
	n=14	n=15

Only one correlation (1:4) is seriously low, and this may be due to the number of scripts rated. Many coefficients are impressively high, which is reassuring. However, this analysis still conflates the different tasks and booklets, and it aggregates the scores given to each separate criterion. To understand better what is going on, we need to look at each criterion separately. For the details of this analysis, see Chapter 10, where we discuss the Writing tasks and the marking criteria.

Mean scores

Correlations do not tell the whole story: two raters may correlate well with each other, and yet give different scores to scripts. One rater may be lenient, the other quite harsh, but if they rank the scripts in the same order, they are likely to correlate highly with each other. It is therefore important also to compare means across raters.

Table 8.3: Mean scores of first and second raters

Variable	Mean	Std Dev	Min	Max	N
R1, Task1	14.36	5.83	.00	30.00	190
R2, Task1	13.08	6.34	.00	31.00	191
R1, Task 2	13.16	7.57	.00	32.00	255
R2, Task 2	12.06	7.26	.00	32.00	240

When the means of different raters are tested statistically for significant differences, the results are as follows:

Table 8.4: Contrasts of means. First and second raters

	Mean	SD	N
Task 1 Rater 1	14.4392	5.752	189
Task 1 Rater 2	13.2222	6.231	189
t = 4.12, df 188, p.0000			
Task 2 Rater 1	13.9328	7.115	238
Task 2 Rater 2	12.1134	7.251	238
t = 6.93 df 237 p=.0000			

The means are significantly different from each other, for both aggregated tasks. This is potentially serious since it means that a student will get a different score, depending upon which rater marks his/ her script. Since different raters rated different candidates, it is not appropriate simply to compare mean scores for each rater, since differences may be due to the difference in ability of candidates. It only makes sense to contrast raters in pairs, marking the same scripts.

Table 8.5: Comparison of mean for pairs of raters, paired t-tests

Paired raters:								
a) 5:10								
		Rater 5		Rater 10				
		Mean	sd	Mean	sd	n	t value	p
	Task 1	10.11	5.66	9.61	6.14	18	.93	.366
	Task 2	10.00	6.69	7.35	5.60	23	3.81	.001
b) 3:5								
		Rater 3		Rater 5				
		Mean	sd	Mean	sd	n	t value	p
	Task 1	14.22	3.95	12.06	4.07	18	5.04	.000
	Task 2	11.70	7.64	9.96	6.37	23	3.87	.001
c) 2:7								
		Rater 2		Rater 7				
		Mean	sd	Mean	sd	n	t value	p
	Task 1	13.78	3.95	14.44	4.81	36	-1.24	.222
	Task 2	13.04	6.42	12.31	6.63	48	1.39	.170
d) 6:9								
		Rater 6		Rater 9				
		Mean	sd	Mean	sd	n	t value	p
	Task 1	12.65	7.00	10.88	5.21	34	2.82	.008
	Task 2	12.35	7.33	10.70	6.38	46	2.72	.009
e) 8:11								
		Rater 8		Rater 11				
		Mean	sd	Mean	sd	n	t value	p
	Task 1	16.9	4.38	13.26	6.22	34	5.69	.000
	Task 2	17.38	6.17	13.82	6.46	39	7.51	.000

f) 3:10								
		Rater 3		Rater 10				
		Mean	sd	Mean	sd	n	t value	p
	Task 1	17.07	5.50	14.07	7.84	15	2.89	.012
	Task 2	15.28	6.35	13.50	9.91	18	1.12	.277

g) 1:4								
		Rater 1		Rater 4				
		Mean	sd	Mean	sd	n	t value	p
	Task 1	15.00	5.73	16.38	8.47	16	-.88	.394
	Task 2	14.80	8.09	12.80	9.65	15	1.53	.149

Clearly the pairs of raters vary in their disagreement. Some pairs show no significant difference in mean scores, whereas others show considerable (and highly significant) differences. It must be emphasised that differences in scores awarded to candidates can lead to unfair results, if single marking is the norm. Double marking is clearly essential in a live examination, as is the training of markers to use suitable rating scales and mark schemes. This is important future work for the Project.

Results of the April Written Pilots

Descriptive statistics

The reader will recall that the tests were compiled into six booklets: two booklets of Listening, and four for Reading/ Use of English. All students took tests of Reading and Use of English, and approximately 500 also took one test of Listening. In all there were 5 Listening tasks with 42 items in total, 13 Reading tasks with 105 items, and 7 Use of English tasks with 80 items. Each item was taken by about 250 students, enabling the calculation of reliable data on item difficulty and discrimination. The mean scores, measures of spread and reliability indices of the tests are given in the following table.

Table 8.6: Descriptive statistics for pilot tests, in raw scores

	List1	List2	Read1	Read2	Read3	Read4
Items (+anchors)	25	27	46	58	80	90
n	244	269	258	253	238	234
Mean	12.8	8.3	24.7	25.5	34.9	25.3
s.d.	4.5	4.7	8.4	10.8	11.4	11.6
Mean %	51	31	54	44	44	28
alpha	.76	.82	.90	.92	.92	.90
mean disc	.38	.42	.43	.43	.37	.33

The standard deviations shown in Table 8.6 are high, meaning that the populations taking the tests were either very heterogenous, or that the tests succeeded in spreading out the population well – which is what tests are usually intended to do. Mean test discriminations were also acceptable, being above the usually recommended .3. Test reliabilities were good, even excellent, except for the first Listening test, which only achieved a low .76. Chapter 12 discusses the various components of this test in more detail.

With the exception of the second Listening booklet and the fourth Reading/ Use of English booklet, the tests appear to have been quite well pitched for this population. Clearly these tests were very difficult, but recall from Chapter 4 (and see Table 8.7 below) that both these tests contained tasks that were intended to be Advanced. These results suggest that that may indeed have been the case.

Table 8.7: Intended difficulty of pilot tests

Listening 1	Anchor plus Basic plus Intermediate
Listening 2	Anchor plus Intermediate plus Advanced
Reading 1	Anchor plus Basic plus Intermediate plus UoE
Reading 2	Anchor plus Basic plus Intermediate plus UoE
Reading 3	Anchor plus 3 Basic plus Intermediate plus UoE
Reading 4	Anchor plus 1 Basic plus 1 Intermediate plus 2 Advanced plus UoE

Test and task difficulty

It is important to remember that the actual students taking one Listening Booklet were different from those taking the other booklet. Similarly, the students taking one Reading/ Use of English booklet were different students from those taking another booklet. Thus differences in mean scores for the different booklets may have simply been due to differences in the ability in the population taking the different booklets. In classical test statistics, the difficulty of an item or a test is a function of the ability of the people taking the test and this makes it impossible to arrive at an estimate of test difficulty that is independent of (i.e. not influenced by) those taking the test. However, recent advances in psychometrics have resulted in the development of Item Response Theory (IRT), which in essence allows us to calculate the difficulty of an item or a test which is independent of the ability of the students who took the test.

Similarly, in classical statistics, the estimate of the ability of test takers is their test score, which is clearly influenced by the difficulty of the test. If students take a difficult test they are likely to get a low score, but that does not necessarily mean that their ability is low: the test may simply have been too difficult. IRT also allows us to calculate a student's ability independent of the difficulty of the test s/he has taken.

Thus we need to calculate IRT estimates of item difficulty and students' ability, independent of each other. This we did using the computer program BigSteps. In addition, however, we need to compare the difficulties of the various booklets. We do this by calibrating all items onto a common scale, using the anchor items.

Anchor tests in Listening (10 items), Reading (10 items) and Use of English (19/20 items) were common to each booklet. Since these anchor items were used across the various test booklets, it was possible to compare each person's score on the anchor items with their score on the items being piloted, and thus to calibrate item difficulty onto a common logit scale. Using this scale, it was then possible to arrive at a calibrated logit score of each person's ability, and thus to arrive at a measure of each individual's ability, regardless of which combination of tests that student had taken. (The anchor tests were developed in a joint project between OKI and CITO, the Dutch National Testing Agency, in 1993-5, as explained in Chapter 4)

The pilot sample was made up of pupils in both Years 10 and 12 (Years 2 and 4 of upper secondary). Just over 1,000 pupils in total took the tests, but a number had to be dropped from the IRT analyses, either because they could not be calibrated (person misfit) or because they did not complete one or other of the tests. We were left with a sample of 944 for the IRT analysis and calibrations.

Since the three subtests – Listening, Reading and Use of English – were intended to measure different constructs, we calibrated the items separately, i.e. we used the

Listening anchors to calibrate the Listening tasks, the Reading anchors to calibrate the Reading tasks and the Use of English anchors to calibrate the Use of English tasks.

The results are presented in Table 8.8 below.

Table 8.8: Mean logit values by task, with intended level

Listening 1 Intended level	Anchor -1.02	Task 2 +0.05 Basic	Task 3 -1.605 Intermediate		
Listening 2 Intended level	Anchor -1.02	Task 2 +1.181 Intermediate	Task 3 +1.921 Advanced		
Reading 1 Intended level	Anchor -0.826	Task 2 -2.298 Basic	Task 3 -0.187 Intermediate		
Reading 2 Intended level	Anchor -0.826	Task 2 -1.64 Basic	Task 3 +0.349 Intermediate		
Reading 3 Intended level	Anchor -0.826	Task 2 -2.624 Basic	Task 3 -2.428 Basic	Task 4 -1.91 Basic	Task 5 +0.71 Intermed
Reading 4 Intended level	Anchor -0.826	Task 2 +0.394 Basic	Task 3 -0.505 Advanced	Task 4 -0.733 Intermed	Task 5 +0.601 Advanced
U of E 1	Anchor1 +1.891	Anchor2 -0.635			
U of E 2 Intended level	Anchor1 +1.891	Anchor2 -0.635	Task 3 +1.404 Advanced		
U of E 3 Intended level	Anchor1 +1.891	Anchor2 -0.668	Task 3 +1.87 Advanced		
U of E 4 Intended level	Anchor1 +1.891	Anchor2 -0.668	Task 3 +1.95 Advanced	Task 4 +0.515 Advanced	Task 5 +1.812 Advanced

The scale may be a little unfamiliar. It is centred around a mean value of zero (0). Negative figures are easy tasks, positive figures are difficult tasks. The range of values in this table is from an easy -2.624 to a fairly difficult +1.95.

It can thus be seen that the item writers' intentions were often wide of the mark: Listening 1 Task 2 was intended to be Basic but is harder than Task 3, intended to be Intermediate. Interestingly, the anchor task, which is said to be at Council of Europe level A2, was easy for these students, suggesting that most Hungarian students would 'pass' the test if it were aimed at A2. In Reading, Task 3 in Reading 4 was intended to be Advanced, yet it was easier than Task 2 in the same booklet, intended to be Basic, and not much more difficult than Task 4, Intermediate. All Use of English tasks are supposed to be at Advanced level, yet there is a range of empirical difficulty.

These results do not invalidate the test tasks per se, but they do emphasise the central importance of pretesting tasks in order to determine their empirical level of difficulty, whatever the intentions of the item writers, and however easy or difficult the tasks may 'appear' to be . We will come back to the issue of level of difficulty in relation to the so-called levels of Basic, Intermediate and Advanced in Chapter 16.

Differences between Year 10 and Year 12 students

Remember that the pilot population was made up of students from both Year 10 and Year 12. For any test task, it is possible to compare the results for the two groups to see whether there are any differences in performance. This can conveniently be done using raw scores, since we are only interested in seeing how individual tasks fared, regardless of other tasks. In other words, we can investigate whether a given task was more suitable for Year 10 students or Year 12 students. This does NOT allow us to say that any given task is 'Basic', i.e. suitable for Year 10 students doing the Basic Examination, since we did not select our populations to represent those who would take the Basic exam, compared with those who would take the Intermediate or Advanced exam. In the absence of clear policies on who will eventually take the Basic exam, we would have no basis for doing so. Nor can we draw firm conclusions from these results about the abilities of those Hungarian school children who study English, because of the pernicious influence of the Rigó utca exams. It is highly likely that our Year 12 population did not include many students who have already passed the Rigó utca exam, as they are unlikely to attend English classes any more. Having said that, our analyses are not invalid, only limited in the conclusions we can draw about the whole population of school students who learn English – which was not our aim in any case. Our aim was to establish and study the levels of the various tasks we had devised, and this we can do.

Table 8.9 gives the mean scores in raw percentages (ie not calibrated data) for all tasks for Years 10 and 12 students separately.

From this table, several points are clear. First, the mean scores for the anchor tasks vary depending on which booklet they were in, that is, which group of students took them. (For example, Listening Anchor in Booklet 1 Year 10 students 42.9%, in Booklet 2 Year 10 students 48.6%, or Reading Anchor Booklet 1, Year 12 students 67.2%, Booklet 3 Year 12 students 54.3%.) This is clear justification for having anchor tasks that enable us to compute their average value regardless of population, and thus to arrive at calibrations for tasks that were taken by different populations. Any pilot test design that administers different tests to different populations MUST have anchor items and MUST calculate IRT calibrations if the results are to be comparable and meaningful.

Table 8.9: Comparison of task difficulty, Year 10 and Year 12

	Year 10	Year 12
	%	%
Listening 1		
Anchor	42.9	54.9
Task 2	24.8	37
Task 3	57.4	62.5
Listening 2		
Anchor	48.6	52.4
Task 2	21.4	20
Task 3	14.9	16.4
Reading1		
Anchor	56	67.2
Task 2	77.8	84.7
Task 3	54.6	53.9
UEAnchor1	23.2	24.4
UEAnchor2	53	64.3
Reading 2		
Anchor	59.5	64.3
Task 2	76.2	75.6
Task 3	43	48.2
UEAnchor1	22	25.3
UEAnchor2	70	72.3
Task 3	30.3	31.7
Reading 3		
Anchor	50.2	54.3
Task 2	77.9	82.5
Task 3	71.1	80
Task 4	63	75
Task 5	20.8	24.7
UEAnchor1	15.6	18.2
UEAnchor2	36	51.8
Task 3	11.3	15.4
Reading 4		
Anchor	47.7	54.1
Task 2	26.4	30.6
Task 3	41.6	43.9
Task 4	45.9	50
Task 5	23.7	27.5
UEAnchor1	12.1	15.1
UEAnchor2	35.2	47
Task 3	7	8.5
Task 4	25.8	27.5
Task 5	7.2	15.1

Secondly, as perhaps expected, Year 12 students usually perform better than Year 10 students, within any one booklet. However, thirdly, sometimes Year 10 students perform somewhat better (Reading Booklet 1, Task 3: Yr 10 54.6%, Yr 12 53.9%; Reading Booklet 2, Task 2: Yr 10 76.2%, Yr 12, 75.6%). And often the differences are so close as to be negligible. These results are important, in that they show that what might be expected to be big differences between Year 10 and Year 12 students are not quite so large. There are several possible reasons for this.

1) The tasks are not good enough to discriminate between these students. In fact, however, as we have seen from the standard deviations and item discrimination figures the tasks do indeed discriminate strong from weak students. So this explanation does not hold.

2) The students in Years 10 and 12 are not so different, because of the way we sampled. However, it will be recalled that in EACH school, we took one Year 10 class, and two Year 12 classes. There is thus NO bias by school or region and it is highly unlikely that the sample of Year 10 classes thus differed systematically from the Year 12 sample, as

they came from exactly the same schools. They were just two years earlier in their English learning career.

3) There ARE no big differences between Year 10 and Year 12 in terms of English achievement. Whilst we will show in Chapter 17 that the overall difference between Year 10 and Year 12 was statistically significant, it may not always have been meaningful. We can see from the results task by task that some tasks were much more successful at distinguishing between Years 10 and 12 than others. The reasons for this will be worth more detailed attention – see Chapters 10-14 on each test paper.

4) However, this also means that we do not have a firm basis in the difference between Years 10 and 12 students for deciding on the level of difficulty (Basic, Intermediate, Advanced) of these pilot tasks. Those levels will have to be determined by other means – see Chapter 16 on standard setting.

Relationship between tests

It is possible to explore the relationship between the various sub-tests of the Written paper in order to see to what extent the various tests overlap, or duplicate each other. If tests overlap, then either they are measuring similar abilities, or one can substitute for the other, even if they measure somewhat different abilities. In general, if sub-tests correlate closely, there is a degree of redundancy, since a score on one test can fairly accurately predict a score on another with which it is correlated.

It is sometime said that Reading and Use of English test are closely related (see, for example, Chapter 13). Table 8.10 explores these relationships.

Table 8.10: Correlation between Reading and Use of English sub-tests

Booklet 1	Booklet 3
n=237	n=224
r=.662	r=.674
Booklet 2	Booklet 4
n=236	n=221
r=.653	r=.650

In fact, the results show that these two tests are reasonably different from each other, sharing roughly 45% of the variance only. These correlations justify having separate tests of Reading and Use of English, since they suggest that they measure somewhat different abilities.

Table 8.11 explores the relationship between listening ability, as measured by either Booklet 1 or 2, and reading or use of English abilities.

Table 8.11: Correlation between listening ability and other abilities

	Reading/UE	Reading alone	UE alone	Reading plus UE
Listening correlation	.657	.577	.570	.631
n size	430	433	439	419

Again, the result shows moderate correlations, but substantial differences, too. Listening appears to be a somewhat different ability from a reading ability or the ability to use English. Separate sub-tests of these abilities are clearly justified.

Table 8.12 explores this relationship in more detail, by comparing performances on each Listening test with each test of Reading and each test of Use of English. The results are remarkably similar to those in Table 8.11.

Table 8.12 Correlation between Listening tests and other tests

	Reading 1	Reading 2	UE 1	UE 2
Listening 1 correlation	.641	.545	.521	.631
n size	105	106	106	109
Listening 2 correlation	.594	.541	.542	.622
n size	108	118	106	123

Finally, it is possible to explore the relationship between each sub-test/measure of ability, and an overall measure of language proficiency (comprising the other two sub-test components).

Table 8.13: Correlation between test sub-tests with total test score

	Reading + UE		Listening + UE		Reading + Listening
Listening r	.631	Reading r	.706	UE r	.701
n size	419	n size	419	n size	419

Table 8.13 shows that each sub-test is indeed substantially related to a more general measure of language ability, suggesting that each sub-test does measure relevant aspects of language proficiency, whilst being substantially distinct from each other sub-test. This provides some evidence for the construct validity of this test battery.