IMPRESSION MARKING ON A GUIDED WRITING TASK

Siriporn Pongsurapipat
Naraporn Rochanachandra
Alan Driver
John Laycock.

Background

1. The CULI EAP (W) course is a one-semester 60-hour 2nd-year course organised on a functional but non-communicative basis. It aims to develop in students the ability to produce a limited range of 'common-core' EAP functions such as exemplification, classification, definition, cause-result, comparision, prediction, as well as to develop paragraphs by adding supplementary detail etc. It is fairly tightly controlled; and since neither content nor communicative purpose are in any way emphasised, students can really only be assessed on a rather limited organisational or, more particularly, broad grammatical basis. (Not surprisingly, teachers find that, in such a controlled course, most students have little difficulty in the 'functional' organizational aspects, but need a great deal of practice in the kinds of structural pattern appropriate to them).
2. The existing marking scheme for each part of the 1982 final exam requires the allocation of marks, in varying ratios, (1) for content + rhetorical aspects according to specified features and (2) for grammatical (including lexical) acceptability and mechanics according to a classification of major and minor errors (not in itself a particularly easy theoretical distinction to make). The scheme-essentially an error-count method-proved fairly reliable and easy to apply for Parts I and II of the paper where the tasks were short, with no question producing an answer of more than about 30 words. It also worked for Part III, since, although the task was slightly longer (around 100 words), all the content (which was entirely non-numerical) was provided in note form, and the functional operations (cause + result, exemplification) required to connect it were relatively simple. However the scheme proved to have a number of unfortunate outcomes for Part IV, which (see Appendix) required students to produce a range of patterns expressing similarity and difference, and manipulate rather difficult, partly numerical data about aircraft in two paragraphs totalling around 200 words. There were 5 principal problems.

1. If the marks were divided equally between content and grammar, in view of the quite easy content/rhetorical demands, but longish answers, grammatically fair and grammatically poor scripts were both obtaining around 50%.
2. More important, the relatively long answers meant that students who had mastered the various ways of expressing similarity and difference taught by the course and were able to produce them correctly were not distinguished from students who could not : both could end up with low scores on the mark allocated for grammar (regardless of whether this was 50% or 30% of the total), because fairly inevitably a large number of grammatical errors were made by most students in handling the unfamiliar numerical data. Many markers felt such a distinction should be shown.
3. The longer the script and the more complex the kinds of error made, the harder it is to decide both the kind and the degree of seriousness of the grammatical error involved. Even the most assiduous marker must eventually be forced to take a 'subjective' decision on what to deduct.
4. If it was done carefully, the marking of this part required on average at least 5 minutes per script.
5. Perhaps as a result of (4) and (5), the reliability produced by this method of marking was extremely low. A random check suggested that in some faculties scripts were receiving up to 25% more marks for this part than they would have if they had been scrutinised in the way that they were in others.

Of these problems, (2) is particularly significant, since it suggests that even in a closely guided task there is to some degree a unitary factor involved which 'does not lend itself easily to componential analysis'. (Oller 1979).

3. Although there exist a number of variations upon error-count marking for example the system described by Oller (1979) of scoring for conformity to correct prose- none of them seemed likely to meet the above problems. The principal other means of marking continuous writing is impression marking.

This is commonly used in free composition writing, where there will of course be a wide range of factors involved : thus Carroll (1980) describes a 'Band 5' (effectively on a scale of 2-9) script as a '*Modest writer.* Conveys basic information competently, but logical structure of presentation will lack clarity. Work will show several slips and formal errors. Use of style and conveyance of tone is present but not consistent. Essay may well lack interest, but the basic message gets through' ; and Mullen (1977) uses a 5-point scale in each of four areas : 'control over English structure, compositional organization, quantity of writing, appropriateness of vocabulary'. Obviously the criteria used in marking must reflect the aims desired for the kind of writing requested ; and since Part IV had few real demands beyond mechanical accuracy, it would not immediately appear that

there should be any need to use any other means of scoring than error-count. However, in his assessment of the error-count method. Heaton (1975) points to two of the problems already listed the inevitable subjective element and the situation where "it is fairly common for an examiner to feel that a composition is worth several marks more or less than the score he has awarded"?; as a result Heaton favours either an impression method, or, where only one marker is available, a graded analytic method, noting that at the elementary level this method could focus mainly on 'grammar and vocabulary'.

Since more than one marker was available, it was decided to conduct a limited experiment in impression marking, but on restricted criteria relevant to the actual question set in Part IV.

Procedure

1. 50 scripts (10 each from five faculties) would be re-marked separately by 4 markers (2 Thai, 2 native-speakers). Markers would give an impression mark based on three criteria :
   1. The inclusion of all the information from the table and its organization into two clearly contrasted paragraphs (using appropriate connecters).
   2. The correct use of appropriate (and varied) patterns for expressing similarity and difference.
   3. General grammatical and mechanical accuracy. Ideally a 3-point scale-Good, Fair, Poor-would be used. (In practice, since the task was very controlled, it was not easy to see a clear enough pattern of divergence, and a 5-point scale proved necessary : G, F+, F, F-, P). No script that used patterns for expressing similarity and difference correctly would receive less than F.
   4. The grades would then be converted into figures and averaged. The actual figures used should give whatever spread was desired. They could, for example 1-3-5-7-9 ; or 4-5-6-7-8. In this case, since the first three questions, marked by the orginal, basically error-count scheme, had produced a wide-spread, and since there were few real criteria other than grammatical ones which could be used for Part IV, which was at the same time likely to invite a fairly large number of mistakes through the problem of students having to handle unfamiliar numerical information, it was felt that a narrower spread would be fair.
   5. Where, on any script, an extreme discrepancy arose (a G and a P), the script should be discussed by the markers concerned. If they could not agree, the script should be put to arbritation. (In practice, new grades were able to be assigned without difficulty).

Results

The results from two faculties are as follows (it should be stressed that for this experiment *no* initial standardisation was carried out).

Faculty 1

	MARKER				MARK			RANK		
St	1	2	3	4	Av4	Av1+2	Old	Av4	Av1+2	Old
A	F-	F-	G	F	6	5	4	=5	=6	7
B	F	G	G	F+	$7\frac{1}{4}$	7	6	3	=3	3
C	G	G	G	F+	$7\frac{3}{4}$	8	$7\frac{3}{4}$	=1	1	1
D	P	P	F	F-	$4\frac{3}{4}$	4	$3\frac{3}{4}$	8	8	8
E	F-	F-	F	P	5	5	$4\frac{1}{4}$	7	=6	=5
F	F	G	F	F	$6\frac{1}{2}$	7	$4\frac{1}{4}$	4	=3	=5
G	P	P	F	P	$4\frac{1}{2}$	4	$5\frac{3}{4}$	.....	.....	.....
H	F+	G	G	G	$7\frac{3}{4}$	$7\frac{1}{2}$	$7\frac{1}{4}$	=1	2	2
I	F	F+	F	F-	6	7	$5\frac{1}{4}$	=5	=3	4
J*	F	F	F+	F-	6	6	7	.....	.....	.....

* Marker 3 orginally gave G, Marker 4, P; they revised these grades on re-reading the script.

Notes

1. Examination of the original marking in Faculty 1 showed that it had been extremely careful. With two exceptions, the impression marking produced a similar rank order, though (according to the scale adopted) with overall higher marks.
2. The impression marking produced a lower score for script J because, while not bad grammatically, it used the same pattern throughout–a limitation the original scheme could not take account of. Script G was incomplete, and the old scheme (probably wrongly) also limited the penalty here. (since both of these scripts, for these reasons, went against the trend of the impression marking according with the original marking, they were excluded from the rank ordering).

Faculty 2

	MARKER				MARK			RANK		
St	1	2	3	4	Av4	Av1+2	Old	Av4	Av1+2	Old
K	F+	G	G	F	7	7½	8½	4	4	3
L	G	G	G	F	7½	8	7½	=2	=1	=7
M	G	G	G	F+	7¾	8	10	1	=1	1
N	F-	P	F	P	4¼	4½	8	10	10	=5
P**	F	F+	F	F-	6	7	7	=7	=5	=9
R	F	F-	F	F-	5½	5½	7½	9	9	=7
S	G	G	G	F	7½	8	8¼	=2	=1	4
T	F-	G	F	F-	6	6½	7	=7	8	=9
U	F	G	F	F-	6¼	7	9	=5	=5	2
V	F	G	F	F-	6¼	7	8	=5	=5	=5

** Marker 2 originally gave G, Marker 4, P; they revised these grades on re-reading the script.

Notes

1. Unlike Faculty 1, the original marking of this question in Faculty 2 was extremely careless: for example, a careful re-marking according to the original scheme of script M gave it a total of 8; of script N, a total of 5¼: but, like script G, script N omitted a great deal of information, and so received a lower score on the impression marking than it could have on the original scheme.
2. The unreliable original marking accounts for the considerable difference in rank order between the original and impression markings.

CONCLUSION

1. There would seem to be grounds for adopting the impression system of marking for questions, like Part IV, that require relatively long answers, even though they are closely guided and have no kind of communicative purpose or 'original' content.
2. Initial standardisation with clear criteria over say 15-20 scripts is important to minimise discrepancies both within and between groups of markers (this would take no longer than the standardisation process for the existing marking scheme).
3. While it has been shown that the consistency between the ratings assigned by judges within a pair will vary between pairs of judges (Mullen 1977), it has also been shown (Heaton 1975) that in impression marking by three or four markers 'the total marks have been found to be far more reliable than the marks awarded by one analytic marker... on the other hand, the marks awarded by one impression marker are less reliable than these awarded by one analytic marker'. This suggests that three or four markers would be ideal. However, as the results show,

except for the inevitable increased bunching, there is little difference between the averages (and resulting rank orders) of all four markers and those of the first two, so that having two markers should present a workable system. Morever, initial standardisation should help to produce greater reliability. (Any discrepancies as great as G–P should continue to be re-marked).

4. The actual marks into which the grades are converted can, if desired, be varied from year to year according to the level of difficulty of the question. But even if a very broad spead is not obtained, the existing marking scheme used for the earlier, shorter parts of the exam should continue to produce quite fine distinctions between students.
5. The advantages of impression marking can be summed up :
   1. It is more reliable.
   2. It takes less time (an estimated 30 seconds per script against at least 5 minutes; even if one marker did look at four scripts, which is probably not necessary, there would still be a saving in time of around 50%).
   3. It enables credit easily to be given for good or desired aspects, while not allowing the negative effect of grammatical or mechanical inaccuracies elsewhere to outweigh them.

REFERENCES

CARROLL, BJ (1980) *Testing Communicative Competence* Oxford : Pergamon Institute of English.

HEATON, JB (1975) *Writing English Language Tests* London : Longman.

MULLEN, K (1977) 'Evaluating Writing Proficiency in ESL' in Brown, HD ; Yorio, CA ; Crymes, RH (ed) *Teaching and Learning English as a Second Language : Trends in Research and Practice* Washington DC : TESOL.

OLLER, JW (1979) *Language Tests at School* London : Longman.

APPENDIX

Part IV from the 1982 Final EAP Writing Test

(Note : the data for the engine types was changed to provide another instance of similarity).

Part IV (10 marks)

1. The incomplete text below comes from a Thai International Airways handout advertising its flights.
2. Use the information in the table below to write two paragraphs describing the similarities and differences between the two types of aircraft to complete the text.
3. Use appropriate ways of expressing similarity and difference.
4. Remember that the information is in note form, and that you may have to add articles or make other changes when you use the information in sentences.

Characteristics	B 747-200 B. Jumbo	DC-10-30
Made in	U.S.A.	U.S.A.
Type of aircraft	wide-bodied jet	wide-bodied jet
Number of seats	371	272
Cargo volume	6200 cubic feet	4618 cubic feet
Length of aircraft	231 feet 4 inches	181 feet 7 inches
Wing span	195 feet 3 inches	135 feet 4 inches
Cruising speed	630 miles per hour	547 miles per hour
Fuel consumption	3275 U.S. gallons per hour	2640 U.S. gallon per hour
Type of engines	General Electric CF6	General Electric CF6
Number of engines	4	3
Used for	long-distance international passenger flights	long-distance international passenger flights

For flights to Europe, Thai Airways International, the national airline of Thailand, uses the B 747-200 B. Jumbo and the DC-10-30. Both kinds of aircraft are widely used commercially nowadays. There are some major similarities between them; however, there are also a number of significant differences between the two kinds of aircraft.

Although there are striking differences between the two kinds of aircraft., they are in world-wide use on international flights since they are both designed to give comfort to passengers on long journeys. So when travelling on them, passengers will not be aware of their differences at all.