LIGHTENING THE ESSAY-MARKING LOAD

Bryan Robinson

MARKING WRITTEN ASSESSMENT ESSAYS is a time-consuming headache for most teachers. Whether you're talking of work written in a foreign language or on a so-called "content" subject - history, literature, or philosophy - in the learner's mother tongue, how do you, the teacher, cope with the difficulties of being objective, fair, and consistent in your marking? And how do you manage to balance these with the ever-present pressures of time?

This question of "inter-examiner" reliability - that is, whether you mark the first batch of essays which you correct when you're feeling fresh and positive on a bright and cheery spring morning, using the same criteria as you employ on the last batch: the ones you start when feeling tired, after dinner on a weekday evening, knowing you're going to miss a good film, and under the added pressure of having - without fail - to produce a set of marks for the learners and your colleagues by nine o'clock the following morning - is a question that has yet to be answered satisfactorily. The aim of this article is to describe a possible solution to this problem - one that worked for me - by moving step-by-step through my recent experience of marking an achievement test of "short" essays on a content-subject.

Although the learners I was working with were not studying English as a Foreign Language, the procedure I intend to describe is - I believe - equally valid. It proposes the writing of a set of "criterion descriptors", labelled by numbers from one to ten, which, once prepared, can be refined and re-used time and time again. I hope to demonstrate that the investment in both time and energy that goes into the initial process bears fruit well into the future, and is of lasting value to both teachers and their learners.

In the course of this article I am going to describe the five-phase process which I followed, explaining how and why I did things, and whether or not I think they worked. In my final discussion I will refer to some of the statistical measures I used in order to try and set my own mind at rest as to my relative success in this task. (To help you follow this, I suggest that before reading further you make sure that you're familiar with the terminology and concepts so ably presented by Graeme Porte in his article earlier in this same issue of Greta. Then, read on.) Finally, I am going to set out in recipe-form a suggested activity for you to use in the classroom in order to start on the road to producing your own sets of criterion descriptors.

Introduction

To begin with I will briefly set out some background information on the learners I was working with, and on the task I had set them. The details referring to the assessment exercise itself - with the exception of the title, obviously - were all established prior to starting the course; the mechanics of marking were not.

- Sample population: 170 first-year, degree course students in the Faculty of Translators and Interpreters.
- Subject: Applied Linguistics for Translation and Interpreting.
- Objective of the exercise: to assess learners grasp of the subject, and their ability to carry out a research exercise using their class notes, the course bibliography, and their knowledge of the characteristics of the genre of text required (SWALES 1992).
- Value of the exercise: 12.5% of the total assessment of the subject.
- Assessment exercise: a short essay (approximately 2 pages A4) answering the question: "What contribution does Applied Linguistics make to the translator's task of overcoming the challenges originated by ambiguity in written texts?"

A "confession"

It would be nice to be able to say that what I am about to describe was a planned process which I had decided on prior to the start of the course. It would be nice to be able to say that... but it wouldn't be true. Like many of you, I guess, I had left the detailed consideration of how to correct these essays until after I had collected them in. Yes, the objectives of the exercise (mentioned above) had already been established and published, and its overall value had more or less been decided, too. But, the most important "detail" was left unclear until the process of marking had already been begun.

Method

Here is my description of the five phases of the process which I undertook.

First phase: Identify your criteria for correcting.

In this section I will go through the first part of the process by which I corrected these essays.

- I started to read the essays, with a pencil in hand, marking both good and bad points appropriately, but without consciously thinking in terms of marks /10, pass or fail, or any other external considerations. Most importantly I tried hard not to compare one essay with another, but rather to be as objective about each of them as possible.
- I read about 15 essays like this and then stopped.
- I next took out a sheet of paper and wrote down a list of the things I had been - (un)consciously? - looking for. These are some of the items on that list:

Answered the question?
Spelling!
Style
Confused
Doesn't understand ideas
No examples
Original examples!
Good use of bibliography
Wide-ranging response

Second phase: Delimit the criteria.

From the list I then tried to join together the different aspects of the essays into the smallest possible number of groups. I wanted to be able to say to myself that each group represented one discrete set of criteria by which I could judge the essays. The final headings for these sets of criteria were:

- *Knowledge of the subject.*
- *Clarity of argument.*
- *Written expression - genre, style, accuracy.*
- *Response to the question.*

I had taken about one and a half hours to reach this point, and decided that it was a good moment to stop. One of the most dangerous parts of correcting any learners' work is over-tiredness. Often we reach the point where we "can't see the wood for the trees", and a ten-minute break can easily remedy that.

The four headings for my embryo "Marking criteria" were now quite clear in my mind, so in order to check their relevance I repeated the first step described above with another batch of ten essays. My criteria held up, and I was now able to think about how I could develop these further.

Third phase: Grade the levels of performance within each set of criteria.

The component areas I had established served the purpose of identifying four separate strands which needed to run throughout the descriptors I now proposed to write. In my own mind I likened what I was about to do to the work of a weaver: here I had four balls of wool, each of a different colour, and I was going to weave a cloth in which at each point you would be fully aware of each one of the component colours, although the blue, say, might predominate here, and the yellow, there. My task began with the need to establish shades of blue or yellow - i.e. numerical grades, ranging between 1 and 10.

Now, putting numbers to descriptors is an extremely difficult process. Carroll & Hall (1985:77) point out the fact that any numerical scale tends to be reduced by two elements as we instinctively shy away from using our maximums and minimums. Consequently, if we propose a scale of five levels, called

- *Advanced*
- *Upper intermediate*
- *Intermediate*
- *Upper Elementary*
- *Elementary*

we really would have three levels to work with, as neither Advanced nor Elementary would tend to be used. In order to give myself space to manoeuvre I decided to use a numerical scale of one to ten, but based on a pairing of numbers at each of five levels:

9 - 10
7 - 8
5 - 6
3 - 4
1 - 2

If I were to follow Carroll & Hall's logic, that would reduce my range to three elements. However, as I needed finally to give each learner a number in order to calculate their overall score, this approach enabled me to work with eight effective levels by using the following approach: When reading an essay, I would initially decide if it was a '3' or a '5'; next I would ask myself "Is this a **good** '3'- almost a '5'- or just a '3'?" A **good** '3', but not a '5' is given a '4'. Is that clear? Technically, this is known as "guided impression marking".

There are two reasons for my using this double-edged numbering system: firstly, this is a way of widening the total number of discrete levels; secondly, it doesn't require more written descriptors. It is a saving on time and energy, but it gains in flexibility.

Here, now, is one of my sets of descriptors broken down into five distinct levels:

Knowledge of the subject.

9 - 10 *Domina los conceptos.*

7 - 8 *Su dominio de los conceptos es amplio aunque no completo.*

5 - 6 *Se perciben algunas lagunas en su dominio del tema.*

3 - 4 *Demuestra ignorancia en aspectos importantes del tema.*

1 - 2 *No demuestra conocimiento de los conceptos más elementales del tema.*

Fourth phase: Produce written descriptors covering all four sets of criteria.

The next phase is really an applied writing exercise for the teacher. I had to weave together my four different-coloured balls of wool into a text that was both clear and concise, and easy to use. The version I decided to use to correct these one hundred and seventy essays was the one that appears on the final page of this article. (If you would like to photocopy this for your own use please do so.)

As you can see, the different strands are woven together in a similar pattern within each of the five discrete levels of descriptors; different shades of each of these appear at each level.

Once my writing task was finished, I went back to the essays and corrected another, different batch of about 20 before having one final look at my set of descriptors. This time I placed the corrected essays in order of the numerical scores I had given them and then went through them, one by one, asking myself whether or not they really were "better" or "worse" than each other. By and large, I was happy with the results of this check, and so I decided to carry on and mark the rest.

Results & Conclusions

Fifth phase: Carry out basic statistical analysis to establish whether or not your results are "normal" (PORTE, op.cit.).

Once all of my correcting was done, I needed to decide whether or not my set of descriptors had produced fair and reliable results for my learners. So, I then carried out some of the statistical analyses which I referred to earlier.

Here are the results of my checks:

MEAN	5.04
MEDIAN	5.0
MAX*	9.0
MIN*	1.0
RANGE	8.0
S.D.	1.72
S.E.	0.61

* 'MAX' refers to the highest score obtained by any one learner in the group. 'MIN' refers to the lowest score.

As you can see, the MEAN and the MEDIAN scores are very similar. This means that the distribution of the marks is very likely to be "normal". This is reinforced by the fact that I was dealing with a very large number of learners (170): the larger the sample, the more likely it is to include instances of all scores, and therefore to be "normal". However, the range of scores for my essay was very high, and this suggests that although the exercise was an achievement test, there were some people who singularly failed to achieve. Consequently the STANDARD DEVIATION (S.D.) is an important number to look at, as this removes distortions caused by unusually high and/or low scores, which the range picks up on.

The S.D. for this group was 1.72. If we look at this in comparison with the range (8.0), we can see that although some individuals scored very high marks MAX = 9.0), and others scored very low marks (MIN 1.0), the majority scored somewhere between the MEAN (5.04) plus the S.D. (1.72), and the MEAN minus the S.D.: that is, within the band from 3.32 to 6.76 (The arithmetic is this: 5.04 - 1.72 = 3.32, 5.04 + 1.72 = 6.76).

So, what conclusions can I reach, then? A small S.D. and a fairly low MEAN initially indicate three alternatives:

- the test I had prepared was too difficult for the students (i.e. I was at fault as tester),
- the students may have been well-prepared, but they didn't do the test very well (i.e. I was at fault as teacher or tester),
- the marking criteria did not differentiate sufficiently well between "better" and "worse" students (i.e. I was at fault as essay-marker).

I am aware that how I reached these possible conclusions may not be immediately apparent to many readers, and it would be useful to go into them in more detail. However, it is not the purpose of this article to do so, and consequently I will skate over them for the moment. Suffice it to say that, by and large these results seem to me to be satisfactory. I expect that the next time I use the same descriptors to mark an essay some learners will perform "better" than they have in this first test because I will have given them a copy of the descriptors before they carry out the exercise: they will know in advance what their - and my - "target" is. Other learners, probably the "more able" in the group, will not find their performance affected: these are the learners who are best able to adapt to whatever type of learning or testing they are confronted with.

At this point, let me point out that with a set of criterion descriptors the kind of analysis involving the "p value"or the "r value" (PORTE op. cit.) is not possible, so to continue my statistical analysis of this particular exercise I will have to look at other measures of correlation - in this case a comparison between the results of my test and that of other content-subject tests carried out by other teachers - to establish which of these conclusions is the most accurate.

However, for the moment, I should say that the third alternative looks the least likely conclusion as the distribution, to which I referred earlier, is quite clearly "normal" in terms of frequency distribution, so it probably comes back to being a teaching deficiency rather than a testing error.

All of this does, of course, raise a very important question with regard to the use of these descriptors: Can they be relied upon? Carrying out further statistical research is obviously time-consuming, so the all-important next step is, always, quite simply to go back to the essays themselves. The teacher always knows best. The experienced teacher's impression is, ninety-nine times out of a hundred, the best measure of learner ability. At this point, with my essays, I returned and looked at what learners had written. I looked at the highest and the lowest scores, and at quite a number of the in-between ones, too. Was I satisfied that these people had achieved more or less the scores that I intuitively expected of them? The answer was "Yes", so that night I slept well!

Closing Comments

In the results presented above there are clearly gaps which need filling. However, within the bounds of this article, and most particularly within the frame of my aims and objectives, these further statistical details are not relevant. Soon I will take the time to correlate these scores with those of other tests in order further to improve this test, as well as to satisfy my own curiosity as to its efficiency.

However, notwithstanding these "gaps", the process I have described here does seem to have worked reasonably well. Let me now review the five phases I have described. They are:

- *Identify your criteria for correcting,*
- *Delimit the criteria,*
- *Grade the levels of performance within each set of criteria,*
- *Produce written descriptors covering all four sets of criteria,*
- *Carry out basic statistical analysis to establish whether or not your results are "normal".*

Following each of these there must be a checking exercise when you return to the essays you have in order to reassess what you've done, and to revise your work accordingly. You must never be afraid to modify descriptors in order to produce new distinctions on the basis of real evidence.

AN EXAMPLE SET OF CRITERION DESCRIPTORS FOR MARKING ASSESSMENT ESSAYS

9 - 10 Capta todos los matices de la pregunta y responde con originalidad y claridad tanto de ideas como de estructura y expresión.
Domina los conceptos.
Maneja con soltura el registro propio del ensayo académico. Los ejemplos y las referencias bibliográficas son pertinentes y están presentados correctamente.

7 - 8 Demuestra haber captado el alcance general de la pregunta y responde con bastante acierto y claridad de argumentación.
Su dominio de los conceptos es amplio aunque no completo.
A veces no es coherente en el uso de las convenciones del género. Utiliza bastantes ejemplos y referencias bibliográficas aunque no siempre cita sus fuentes adecuadamente.

5 - 6 Parece haber entendido el alcance general de la pregunta y la respuesta es adecuada aunque un tanto limitada. A veces la argumentación es confusa.
Se perciben algunas lagunas en su dominio del tema.
Su uso de las normas que caracterizan el ensayo académico no es siempre correcto. Cita algunos ejemplos y/o referencias bibliográficas aunque no siempre de forma apropiada.
Hay algunos "lapsus" de ortografía y/o puntuación que no afectan a la comprensión del trabajo pero sí le restan calidad.

3 - 4 No ha demostrado haber entendido el alcance global de la pregunta. La respuesta es inadecuada y a veces confusa.
Demuestra ignorancia en aspectos importantes del tema.
Demuestra una falta de conocimiento del género. Apenas utiliza ejemplos y/o referencias bibliográficas. Tampoco son siempre acertados los que utiliza.
Existen errores frecuentes de ortografía y/o puntuación.

1 - 2 No ha entendido la pregunta.
No demuestra conocimiento de los conceptos más elementales del tema.
No conoce el género de texto. No utiliza ejemplos ni referencias bibliográficas.
El texto está plagado de errores lingüístico-técnicos.

D-I-Y Descriptor-writing

WRITING YOUR OWN DESCRIPTORS for essay-marking is not that easy a task. This activity (which mirrors the process described above) is designed to enable you and your learners to collaborate in the process - and share the work-load around!

Level: Secondary school.

Timing: 3 classes of 45-50 minutes each.

1. In the first class, set your learners an essay to write. Tell them that they are going to mark it themselves in successive classes. The essay can be on any topic at all -they could choose it themselves, as long as they all write about the same subject. Collect in the essays.
2. In the second class distribute the essays, one to each learner, making sure that nobody gets their own work.
3. Ask them to spend ten minutes "correcting" the essays in pencil. Tell them they must mark an equal number of positive and negative aspects.
4. The class then divides into subgroups of a maximum of five people each. Again, take care that nobody is going to work on their own essay.
5. Each sub-group puts together the essays they have corrected and makes a list of the points they have been looking at. A typical list would include things like content, spelling, grammar, and vocabulary.
6. They should reduce this list to the minimum number of discrete ideas, and then go back to the essays they have to check they haven't left anything out.
7. All of the groups now put their ideas together and you guide them towards one single group of criteria headings about which they all agree. At the end of this class collect in the essays again.
8. In the third class, each of the criteria areas needs to be graded in terms of levels of performance. Divide the class into as many groups as there are criteria areas. If the groups are very large subdivide them.
9. Each group is going to work on one specific area. Ask the learners to think in terms of three levels of performance: Advanced, Intermediate, and Elementary. For their allotted area they should write a description of each level.
10. Distribute essays to each group and ask them to grade each according to their descriptors. Then they can revise the descriptors if necessary.
11. Form new groups so that one person from each of the criteria areas is now in each. First, each presents the work their group has produced. Then, together they weave together the different strands to come up with a set of descriptors for each of the three levels.
12. The groups go back to the essays and mark them using their composite sets of descriptors, and then revise them once again if needed.
13. Collect in the descriptors and you now have a model from which you can go on to develop your own, more sophisticated, three-, five-, or whatever level version.
14. In later classes, use the different descriptor the learners have produced for them to correct their own and each other's work. Now, both you and they know what the targets they are expected to achieve should be.

Bryan Robinson es profesor asociado del Departamento de Lingüística Aplicada a la Traducción e Interpretación de la Universidad de Granada y responsable de exámenes en Inglés como lengua extranjera del Bachillerato Internacional.

Bibliography

Carroll, Brendan J.; Hall, P.J.: Make Your Own Language Tests. Oxford: Pergamon Institute of English, 1985.

Porte, G.: "Testing your test - Basic methods for analyzing the efficiency of your tests", in *Greta,* Vol. 1 Nº 2. Granada: Greta, Asociación de Profesores de Inglés, 1994.

Acknowledgements

My thanks are due to Brendan Carroll and other colleagues of the International Baccalaureate Language B Pilot Project Standing Committee. Many of the discussions we have had have greatly aided me in formulating the principles behind the process described in this article. Thanks also to my colleague Marián Hens for her useful comments on a final draft of this article.