

# PITFALLS IN THE SUBJECTIVE SCORING OF SPEAKING AND WRITING - AND SOME WAYS ROUND THEM

Alastair Ker  
Language Institute  
University of Waikato

## INTRODUCTION

Much as we may detest it, many of us are called upon constantly to quantify our students' ability or achievement in terms of marks or grades. This process is especially onerous when the score relies heavily on our judgement - in the terminology of testing, the assessment is "subjective". We should perhaps spare a thought for our distant relatives, the judges of Olympic freestyle ice skating routines, who, no doubt with well-established guidelines, still have to take in a performance in the round and hold up their score cards in a more or less instantaneous decision about whether what they've just seen was a 9.5 or perhaps just a 9. We, at least, generally have the benefit of some thinking time and the opportunity to involve and consult colleagues and students. Yet there is a sense in which what both the ice skating adjudicator and the language teacher using subjective scoring are doing - using personal judgement to assign numbers to a complex human activity - seems, like the gravity-defying feats of Olympic ice skaters, both impressive and impossible at the same time. It is no wonder, then, that subjective assessment provokes a great deal of critical reflection by those of us who use it. This article is one product of that reflection. In it I will trace some of the causes of the unease which subjective scoring provokes, and examine ways in which those factors can be minimised.

## OBJECTIVE AND SUBJECTIVE ASSESSMENT

Although it is of course common for subjective assessment to take place without any scoring occurring, and this is, in fact a situation with which some teachers feel more comfortable, I use "assessment" here to refer to assessments in which students receive some kind of mark or grade.

### Objective assessment

For some decades now, the psychometric tendency in language testing, which has sought to place it on a scientific basis, has put a high premium on reliable measurement. The mode of assessment which fits its paradigm best is "objective" assessment - for example, in multiple choice tests and others in which there is a correct answer to every question, since the method on which test scores are calculated is explicit and fully replicable.

### The rise of subjective assessment

Objective methods of assessment are and will always be an important place in language testing, in spite of the questions about validity just discussed. More recently, however, the ascendancy of Communicative Language Teaching has required types of assessment which

test what CLT seeks to promote - namely "communicative competence" (and its manifestation in "communicative performance"). The principles behind communicative performance testing, in particular, give preference to communication in "authentic" situations where there is no predictable outcome. This use of language in the round is much more problematic to score than a set of test items with a limited range of possible answers.

While these changes in what we are expected to assess are most obvious in the testing of oral skills, a parallel situation exists with written tasks. In writing, features like organisation of discourse and sense of audience which contribute at a macroscopic level to the success of the writing as communication are now given equal or greater prominence in marking schemes to that which grammatical accuracy, punctuation and other "mechanics" receive. For both oral and written skills then, new demands have been placed on assessment (cf. Oskarsson, 1981, pp. 225-226), and subjective methods have been looked to to fulfil this role. In doing so, they complement objective methods of assessment.

### **The relationship between subjective and objective methods of assessment**

We can sum up the relationship between subjective and objective assessment by using another sporting metaphor. Subjective assessment is to objective assessment as orienteering is to road running. Where the tarseal road of the exhaustive answer key ends, the assessor is forced to become more self-reliant. Orienteers have to find their own way through difficult, often forested country without clear-cut paths, with map and compass. Like any human being without clear landmarks for orientation, orienteers are prone to lose their way, though perhaps less often than the rest of us! People lost in the desert, for example, apparently tend to end up walking in a circle to the right. This kind of in-built tendency to err is also a danger for subjective assessment, as we shall see shortly.

In other words, subjective assessment typically operates in areas in which there is no substitute for human judgement. Nevertheless, for that very reason, it has a number of in-built pitfalls.

## **PITFALLS OF SUBJECTIVE ASSESSMENT**

### **Factors influencing markers' judgement**

The first type of pitfall relates to the myriad of ways in which our judgement can be influenced by extraneous or secondary considerations. These are generally well known, but here are some examples. If we already know the students whose work we are assessing, there is room for the "halo" effect, whereby a student receives a score based on our existing perception of their ability. For an example, refer to Reves (1991). It is also possible for our judgement of a person's performance to be affected by our rapport with them (or lack of it) and by circumstances as peripheral to the assessment as appearance and demeanour (in an oral interview or classroom work) or, in the case of written work, by handwriting and whether an assignment has been typed or not. It is natural to deny that *we* are influenced in this way, but it is perhaps safer to assume that we are.

Another human quality which can make subjective assessment subjective in the negative sense of the word is our preference for judging by comparison. Even when we have explicit guidelines for making our assessment decisions, our sense of what is possible or desirable in

a particular assessment task is frequently moulded by the performance of the best students. In my experience, when a good seminar presentation, for example, follows a mediocre one, the strengths of the latter will be highlighted; when it follows an excellent one, its weaknesses will tend to stand out. (This is not unlike the optical illusions in which a line can be made to look longer or shorter according to the direction of the arrows placed at its ends.)

Any influence that might lead the same person to assess the same piece of work differently under different circumstances threatens what is known technically as "intra-rater reliability". Changes in mood, tiredness and numbness from excessive marking are other familiar factors which affect intra-rater reliability.

### **Double marking - Are two heads better than one?**

So far I have given examples of how a single marker's assessment can be influenced by extraneous factors. These phenomena are well-known, and for this reason, double marking (here used to refer to marking by two or more independent markers) is universally regarded as desirable in subjective assessments, even if for practical reasons it is not always convenient to implement and is often not used, even by examination boards (see Alderson and Buck, 1993, for the results of a recent survey which the authors carried out in Britain). Double marking offers an opportunity for markers to supplement each other's views of a student's performance and, in the more straightforward cases, to reach a consensus if there has been some disagreement.

However, it needs to be acknowledged that markers do not agree in every case, especially if the assessment procedure is not explicit. Such disagreement is healthy, because it can contribute to a more rounded picture of the work being assessed. The discussion which it creates challenges markers to examine their marking criteria. At the same time, it is important for the students whose work is being assessed that a consistent procedure is developed for dealing with such disagreements, i.e. a moderation procedure which allows markers to agree to differ, but which still ensures that a student's work is given the fairest possible consideration.

## **HOLISTIC AND ANALYTIC SCORING**

When you are faced with the challenge of scoring something as complex as a student's communicative competence (or "proficiency" or "ability") in either speaking or writing, there are two general approaches open to you. One is to attempt an assessment of what you hear or read as a whole and assign it a score on a single scale (holistic scoring). The other is to assign scores to its components individually (analytic scoring). In fact, most forms of subjective scoring attempt a synthesis of the two approaches.

### **Holistic scoring**

Holistic scoring has its theoretical justification in the familiar dictum that the whole is often greater than (or less than) the sum of its parts and its practical justification in its rapidity. It can also produce final scores with satisfactory or even high levels of reliability. The pre-conditions for this reliability are thorough training of markers based on clear procedures for assessment e.g. an assessment scale together with a well-constructed set of descriptors stating

the criteria for awarding each step in the scale; double marking; and a moderation procedure in place in case of disagreement.

In her revealing study entitled "Holistic assessment: what goes on in the rater's mind?", Caroline Vaughan asks whether this is sufficient (Vaughan, 1991). Though acknowledging that holistic assessment has a place, she questions whether it is reliable enough to be used as a basis for important assessment decisions, and asks just how holistic it really is. In her research she used a think-aloud procedure to get teachers to verbalise the criteria they were using to rate student essays on a six-point holistic scale. The teachers all had experience of marking essays to the scale, for which band descriptors were provided.

What the think-aloud procedures showed, though, was that many of the raters effectively ignored the criteria they had been given and used their own engrained criteria instead. These differed from one rater to the next. For example, two of the nine teachers in the study appeared to be dominated by first impressions; others by other single factors, such as grammar; and others by varying combinations of factors. The raters differed both in the factors they took into account and the weighting they gave to them. They differed, too, in their policy on how to deal with borderline cases. One rater made it clear that any borderline student would be given the benefit of the doubt, while another is quoted as saying "The first thing I do is I look for things that would make it not a passing essay" (Vaughan, 1991, p. 118).

Vaughan points out that the six essays which the teachers had been given to rate were all ones which previous marking identified as borderline, and which may therefore have been inherently difficult to categorise. This may have encouraged the raters to revert to their accustomed highly personal rating procedures. But this is, I think, a charitable interpretation. I suspect that most of us would recognise ourselves somewhere in the gallery of rater profiles which Vaughan presents us with. We all have our own slightly different expectations from a good piece of writing, role play or seminar presentation, as well as pet likes and dislikes, and it is very difficult to switch these off when making assessments.

But is Vaughan's study an indictment of holistic assessment in general? Some would argue that the situation which Vaughan portrays is due to the inadequacy not of holistic scoring as such but rather of the training which the raters had received, and their failure to internalise the procedure which had been laid down. Others would argue that the problem does indeed lie with holistic scoring, in so far as it leaves the raters without any mechanism for assessing and weighing up the relative worth of the various components of a student's performance, especially in the case where the student is strong in some areas and weak in others and therefore does not fit the assumption behind the performance descriptors. They see analytic assessment as a more explicit means of assessment which results in a profile of the learner's ability and which is thus able to accommodate and indeed to highlight such discrepancies when they occur.

### **Analytic scoring**

There can be little doubt that analytic marking schemes have the *potential* to achieve greater explicitness of marking and to provide a better basis for feedback to the learner, but here, too, there are some important decisions to be made. What aspects of language use should be assessed? And if the students need to be given an overall score at the end, what relative weighting should each of these receive? One tendency with analytic assessment is for the

assessment procedure to become so clumsy that it is difficult to apply, so that some categories are not assessed properly (in the case of real-time assessment) or the marking scheme becomes very time-consuming to apply.

A further, related problem results from the frequent inclusion of categories such as socio-linguistic appropriacy and use of communication strategies in analytic schemes for assessing speaking ability (owing to the influence of the communicative competence concept). In my experience, it is very difficult in a real-time assessment to assess these features of language use independently of one's overall impression of the student's success in communication. Whereas the latter quality lends itself to being assessed on a multi-point scale, both socio-linguistic appropriacy and strategy use can be difficult to pin down, particularly since, if they are used skilfully, the means by which they are achieved may be virtually invisible, and therefore difficult to grade explicitly.

### ASSESSMENT SCALES

Clearly most of the "problems" alluded to above can be minimised by adapting the assessment scheme to fit the task more closely. Once again, there is a lot to consider. In most assessments there are items which are better treated by using yes/no boxes than in applying a scale. Supposing a scale is to be used, there is an inevitable tension about whether it should have many steps to allow differentiation or as few as possible to make it simpler to apply. The answer to this depends partly on the purpose of the scale. A scale may be used, of course, just to create some kind of ranking between the people being assessed, or it may be bound in to some form of criterion referencing. In the former case, the scale needs to contain sufficient steps to achieve the desired degree of ranking. In the latter case, Weir (1990) suggests that the scale needs to be as simple as possible, to enable descriptors to be written which can be distinguished from one another and which will mean the same thing to different people.

Weir (1991), Hamp-Lyons (1991) and Upshur and Turner (1995) all advocate basing the construction of assessment scales and criteria on the assessment tasks themselves to ensure that the qualities which are being assessed are salient ones for that task and that the criteria for different levels of performance do enable the assessor to distinguish these clearly. Upshur and Turner suggest replacing conventional scales with what they call "empirically-derived, binary choice, boundary-definition scales" (p. 6). These are a kind of decision tree which allows the marker to assign a grade by answering a short series of yes/no questions about the work they are assessing. The questions to be asked are developed from a group of sample performances which have been ranked on the basis of impression marking. The binary decision-making involved makes this kind of procedure very reliable, according to the authors. One reason for this is that choosing between two clearcut alternatives is the kind of decision making we human beings find easiest. This method holds a lot of promise for situations where the qualities of the work being assessed cluster together in the way reflected in the structure of the decision tree. However, in my experience with written assignments, students' work often shows much more complex patterns of strength and weakness which do not fit readily such a simple template. These "difficult cases", which cause problems both for conventional scales and for alternative ones like Upshur and Turner's, are discussed in the next section.

## DIFFICULT CASES

### The Law of the Ill-Fitting Scale

Much of the unease that teachers feel about subjective scoring can be laid at the door of what might be called the Law of the Ill-Fitting Scale - no matter how well-defined the steps of a scale are, there are always candidates who the descriptors fail to describe. This may be because their performance simply falls between these points, or because it has (or lacks) qualities which the scale simply fails to capture. Spolsky (1995, pp. 349 ff.) writes in this context of the futility of the quest for "the Holy Scale". Any scale, therefore, needs to be accompanied by instructions on what to do when it does not fit.

### Procedures for dealing with difficult cases

Whatever the reason for a case being difficult or borderline, we are not going to achieve consistency in our scoring (either personally or amongst scorers) unless we have an explicit procedure for dealing with them. If second marking and moderation are in place, the matter may be resolved in the course of that process. Another strategy to which markers sometimes resort is to create half-steps between our original scale steps. This occurs quite often in universities, where essays are still usually marked on an A - E scale. The scale already has official part-steps built in, since work can be given + or - grades, as well as the straight grades themselves. Nevertheless, students are sometimes awarded grades that are borderline between these as well.

The most common example of grade sub-division is the B++ which is awarded for very good work at the top of the B+ range which is not good enough, in the marker's eye, to be given an A-. Given that the psychological difference between a grade in the A range (albeit an A-) and a grade in the B range is great, this extra step may be justified. But because university teachers much less often create half-steps between, say, a B and a B+, it could be argued that students whose work is on the border between *these* grades do not receive the same degree of consideration.

Another way of dealing with truly borderline cases is to use a rounding principle: "If in doubt, award the higher of the two grades" or "the lesser of the two grades". As we saw from the examples quoted by Vaughan, it is possible for two markers working alongside each other and to the same criteria each to be applying opposite principles. This can make assessment dangerously like a lucky dip, and surely provides a strong argument for markers agreeing not only on criteria for applying a scale, but also on a consistent procedure for dealing with borderline cases.

### Glass ceilings and glass floors

The choice of procedure is not an arbitrary one, especially if there is already an uneven distribution of marks along a marking scale. This may reflect the quality of the assessed work relative to the criteria set for the award of particular scores or grades. But it may also suggest that, in the marker's head at least, a new scale has been defined which only uses part of the scope available in the original one. This may take the form of a "glass ceiling" effect, whereby certain grades are attainable in theory but very seldom given in practice, or of a "glass floor", whereby even a very weak performance is given the minimum pass grade rather

than a fail grade. In the process, a "scale within a scale" is created. This carries with it the danger that the resulting procedure will be a hybrid which is not ideally suited to its purpose.

## DESIGNING OPTIMAL PROCEDURES FOR SUBJECTIVE SCORING

The previous discussion set out to demonstrate the complexity of subjective scoring and the potential for inconsistency and unfairness which it contains. The dilemma we face as practitioners is that, in full awareness of these pitfalls, we still have to continue designing and using assessment procedures as part of our work - we do not have the luxury of descending into a Hamlet-like state of inaction. What principles then will help us choose the least imperfect scoring procedures for subjective assessments? How can they be implemented so as to minimise the numerous threats to the reliability and validity of any assessment? The following list is not exhaustive, but it may be helpful:

### The individual teacher

#### *Training*

Testing and assessment are sometimes seen as esoteric subjects which can be left up to the staff testing expert, or, at the other extreme, as something anyone can do. To accept this is a recipe for leaving things as they are. If we wish to be professional in our conduct of assessment, then each individual teacher must use the available opportunities both within the institutions where they train and where they work for training in testing and assessment. Professional bodies like TESOLANZ and its regional affiliates have a useful role to play in organising workshops which deal with particular aspects of assessment (for advice on how to run a workshop on speaking assessment, refer to Knight, 1992).

#### *Attitudes*

It is a pre-condition for sound subjective scoring that teachers and assessors in general have a healthy mixture of self-respect and critical self-awareness, as well as a willingness to work continuously on refining their assessment procedures. These are not qualities which one either does or does not have - their development depends as well on the value which the institutions in which we work place on good assessment practices.

### Scoring approaches and procedures

#### *Choice of scoring approach*

Often the endpoint of a scoring procedure is beyond our control - i.e., we are expected to express the result of an assessment as, say, a grade from A-E or a mark out of twenty. However, as in orienteering, there is a whole variety of routes by which the endpoint can be reached. In choosing the route which is best suited to a particular assessment task, it may be helpful to consider the following questions:

- Are you interested in scoring a student's performance as a whole, or only in very specific aspects of it?
- Is there a pre-specified format in which the outcome of the scoring process is to be reported? If not, what format would give the student and other people involved the most useful information about the performance of the assessed task?

- If a final global mark or grade has to be given, should the scoring of individual aspects of the student's performance be binding in calculating the final score, or should it be treated merely as a guide?
- If the global score is to be the sum of the individual scores, should these all be given an equal weighting, or is there a reason to weight some more heavily than others? (Hamp-Lyons, 1991, p. 249 recommends an equal weighting unless there is a good overriding reason for doing otherwise.)

### ***Simplicity***

Use the simplest procedure possible which does justice to the complexity of what you are assessing. The reason for preferring simplicity is partly to be economical with time, but also to maximise your control of the procedure. A procedure is there to support assessors and to help them articulate their judgement, and the more complex it becomes, the more likely it is to turn into an obstacle race instead.

### ***Specificity of descriptors***

Make criteria for awarding a particular mark or grade as specific as possible to enable consistent decisions to be made. Catchall expressions like "some errors in pronunciation" are meaningless because they cover everything from a few to a lot. For sets of descriptors which might be helpful models, see Hamp-Lyons (1991) and Weir (1990).

### ***Tailor-made procedures***

The best scoring procedures will ultimately be those which have been designed and refined specifically to fit the assessment task itself. Off-the-shelf scales and scoring rubrics are a good source of ideas, but are unlikely to be transferable unaltered to your particular purpose. The development of the assessment task and of the scoring procedure should have a reciprocal effect on one another. One way to help achieve this is to use the guidelines which you give your students on how to approach the assessment task as headings on your scoring and comment sheet, and then refine the content of each in the light of your experience with the assessment task.

### ***Modifying scoring procedures***

If a particular category in a scoring procedure is difficult to score or seldom produces useful or telling information, be prepared to change the format or the descriptors, to replace it with something more tightly defined or even to eliminate it altogether if necessary.

### **Ensuring reliability**

All of the suggestions made already should in themselves contribute to making subjective scoring more reliable. The desirability of supplementary measures such as double marking and moderation has already been discussed, along with the need for uniform procedures for dealing with borderline or otherwise difficult cases.

For recurring, large-scale assessment tasks, a collection of sample scripts (for oral work, tapes or video clips) is essential as a basis for marker training. Hamp-Lyons (1991, p. 274) points out that several samples are likely to be needed at each level, because different students will be awarded the same grade for different reasons.

## CONCLUSION

### **An eye for the surprising**

Finally, it goes almost without saying that as teachers we should be open to novel or unexpected approaches by students to assessment tasks, and be prepared to give them due acknowledgement. Students who try something different can be penalised for it, even by well-intentioned assessors, if what they do is not foreseen in the assessment procedure. Dancing a bolero in a ballroom dancing competition should not lead to instant disqualification, in other words. Subjective scoring lives from the vivacity of human judgement, and from its ability to do justice, among other things, to the vivacity of our students and their language use in a way that no mechanical procedure can.

## REFERENCES

Alderson, J.C. & Buck, G. (1993). Standards in testing: a study of the practice of UK examination boards in EFL/ESL testing. *Language Testing*, 10, 1-23.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood, NJ: Ablex.

Knight, B. (1992). Assessing speaking skills: a workshop for teacher development. *ELT Journal*, 46, 294-302.

Oskarsson, M. (1981). Subjective and objective assessment of foreign language performance. In J.A.S. Read (Ed.), *Directions in language testing* (pp. 225-239). Singapore: Singapore University Press.

Reves, T. (1991). From testing research to educational policy: A comprehensive test of oral proficiency. In J.C. Alderson & B. North (Eds.), *Language testing in the 1990s*. London: Macmillan.

Spolsky, B. (1995). *Measured words*. Oxford: Oxford University Press.

Upshur, J. & Turner, C. (1995). Constructing rating scales for second language tests. *ELT Journal*, 49, 3-12.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex.

Weir, C. (1990). *Communicative language testing*. Hemel Hempstead: Prentice-Hall.