

Chapter 2

Backwards Assessment Explanations: Implications for Teaching and Assessment Practice

D. Royce Sadler

Abstract A common approach in assessing student achievement is to advise students before the event of the criteria that will be used in judging their levels of performance. There are strong grounds for being wary about that approach. The main reason is the impossibility of accurately anticipating all the criteria that will turn out to be important in arriving at a sound judgment before actually scrutinising individual student works, which are all more or less different from one another. If, instead, a judgment is made first, taking on board whatever aspects of the work or performance seem to be relevant to the judgment being made, the intellectual processes involved in scrutinising the work are radically changed. Not only that, but the explanations for the judgments will not only differ from one another but also be more tailored to each student work or performance. If students can themselves be inducted into this type of thinking and practice, they will become better able to monitor the quality of their own work while it is under preparation and production.

Keywords Assessment criteria • Creative and performing arts • Assessment rubrics • Criteria and standards • Criteria based assessment • Teacher feedback

2.1 Introduction

The title of this chapter may initially seem curious and perhaps a little awkward, so some explaining is in order. The three key words are *backwards*, *assessment* and *explanations*. Starting with the middle one, this chapter is about assessment, predominantly as it is practiced in many taught courses in higher education institutions across a number of countries. Some of the material is applicable to other levels of education as well, but they are not the first priority. As in all areas of education,

This chapter is based on a Keynote Address to the Assessment in Music Conference held at the Queensland Conservatorium, Griffith University, Brisbane on Tuesday 16 July 2013.

D.R. Sadler (✉)

School of Education, The University of Queensland, St Lucia, Brisbane, QLD, 4072, Australia

e-mail: d.sadler@uq.edu.au

assessment practice has been steadily transformed over recent decades, sometimes with insufficient scrutiny of the conceptual or philosophical bases of developments and a lack of sensitivity to side effects. The same could be said of older ways of doing things as well. It is now possible to look back at some of the newer practices with the benefit of hindsight, reflect on what is happening and see whether and why some rethinking might be necessary. That is what this chapter is about.

However, the “backwards” in the title does not refer to retrospectivity alone—looking back, say, at the changes in assessment practice over the last years. Backwards also refers to two particular aspects of assessment practice which, if a key characteristic is reversed from the orthodox direction, can lead to improved learning. These two aspects have been identified by a close examination of certain pedagogical processes and authority relationships evident in interactions between teachers and learners. The third term, explanations, is in the title because constructing explanations for why certain decisions or judgments have been made or why certain events occurred requires focused thinking. The act of formulating and expressing an explanation in words changes the thought patterns of the communicator, usually by way of increased precision. It also produces a communication which can be analysed, challenged (if necessary) and refined so that it leads to changes in the ways other people think about things. Articulation provides a tool for consolidating understanding and learning.

2.2 Terminology

Some common terms used in discussions about educational assessment have a variety of meanings (or just nuances), so a few of these are now clarified for purposes of this chapter. The first of these is learning. A person will be said to have learned something when they can do, or now know, something they were unable to do or did not know before. In general, the person should be capable of demonstrating their learning by being able to perform a relevant task on demand (that is, whenever called for), independently of particular others, and to a satisfactory level. Learning results in a person’s increased or improved capability, and their learning is something they “possess” and carry around with them. Capability normally decays over time—sometimes slowly, sometimes quickly—but if circumstances are conducive, it may be regenerated or restored to the full level previously held without huge cost or effort. “Independently of particular others” means that a person’s performance can be demonstrated without support, coaching, scaffolding, prompting or editing by the teacher or another student. The person can operate in a self-determining way. Of course, in many real-life activities, people actually perform in their field of expertise in the company of and with the contributions of others, as in bands, ensembles, orchestras, film and screen productions, sporting teams and industry work groups. The term *achievement* follows on from and expands the third element in the definition of learning—“to a satisfactory level”. Achievement denotes a certain level reached, accomplished or attained

successfully, or a task brought to a successful end. High-level achievement is usually attained through exertion, effort, skill, perseverance and practice.

Learning and achievement are fundamental concepts in education, and it might be thought that there is little need to labour the point. That would be true except for the way achievement is being assessed in many higher education contexts. The pure concept of achievement has become corrupted by the practice of adding into a total mark, score or grade a variety of elements that are not part of achievement itself. This includes credits for participation in various learning activities (such as class discussions, contributions to online forums, and journals or logs of activities engaged in). These may all assist learning to occur, but they themselves are not the learning. Similarly, demerits or penalties for occurrences of non-achievements (such as late arrival or late submission of a piece of work) may reduce what purports to be a *measure* of learning, but do not reduce the learning itself. It is, of course, important for students to be punctual, cooperative and diligent, but adjusting the achievement score or grade is not the way to do it. Other incentives must be used. In any case, it is not uncommon for certain patterns of behaviour to change rapidly when the stakes are high. For example, a musician who misses out on a critical gig because of lateness quickly gets the message. Punctuality is simply a volitional behaviour which can be changed at will; it is not part of achievement. Assessments of achievement should reflect the “status” or “level” reached, regardless of the process, learning pathway, or time taken to acquire that level. The common practice of adding marks or scores during a learning period (such as a semester) distorts what should be an assessment purely of the level of achievement reached by the end of the learning period. This issue is explored thoroughly in Sadler (2010a). In passing, note that the concept of “achievement” may go by other names in different contexts, including attainment, accomplishment, capability, competence (but not competencies or competences), or performance (but not “a” performance). Finally, in this chapter, “appraisal” and “judgment” are used interchangeably.

2.3 Preset Criteria, Rubrics and Scoring Sheets

The first of the two aspects of assessment practice for which reversal is needed is the practice of specifying the criteria for appraisal at the same time students are given an assessment task. The special interest is in assessment tasks that require students to construct complex student works or performances. Appraising the quality of such responses typically involves multiple criteria, often interlocking or overlapping. The assessor’s brain is the main “instrument” for what is clearly a “qualitative” judgment, meaning that it involves no quantification—aspects that are counted or measured. There are no formal rules or routines which would allow a person new to a particular type of response to make consistently sound judgments. Judgments are also involved when criteria are specified in advance; the appraiser needs to have knowledge of what the criteria mean and imply, and how to apply them to works or performances of the relevant type and in a given context. Criteria

are usually formulated and presented in the form of rubrics (cross-tabulations of criteria and standards); grade descriptors; or detailed scoring schemes. The rationale for advising students before they begin working on their responses is quite straightforward. It is to achieve greater transparency for students (no surprises sprung on them after the event); improvements in the quality of works students submit (because they can be guided by the criteria during their production processes); and greater objectivity in marking. In short, the aim is to make the assessment experience fairer, more open and more ethical.

Few would disagree with the spirit of these intentions and what may appear to be persuasive logic. Yet despite those, there are strong grounds for being wary about deciding criteria and standards first and then applying them meticulously to arrive at a judgment. But if the order is to be changed, what happens to fairness and transparency, and what could substitute for using explicit criteria to guide production to create high quality works? Another initial concern could be that an alternative might be more labour intensive for teachers and markers, or that the change would precipitate appeals from students that their work was judged unfairly because they were in the dark about the criteria to be used. Criteria-based marking is not universal practice of course; holistic judgments are common in many areas of the creative and performing arts, but there has been a trend in recent years in a substantial number of institutions to require, as part of their assessment policies for all assessments, that explicit criteria and standards be stated in advance.

Mention has been made of improving the objectivity of appraisal. The issue of subjectivity-objectivity has been hotly debated in philosophical and academic circles for a very long time. It has an enormous literature, complicated by the fact that the terms do not have the same meanings in different fields. Technically, some would classify a judgment by an individual in relative isolation as “subjective” by definition. For some reason, it is often spuriously assumed that a judgment reached through quantifiable processes makes them objective. A particularly useful contribution to this issue is a chapter by Scriven (1972), even though it was not written specifically in relation to assessing student achievement. Scriven pointed out that the quality of a judgment made by a single assessor is not automatically suspect and deserving of dismissal merely because it was made without collaboration and without the help of formal algorithms or instrumentation. Many professionals in the practice of their professions constantly have to rely on their own professional judgments as well as those of others from time to time. Many of their decisions are not independently verified by accurate measurements using a standardised laboratory tests, even though it may be possible to carry out such tests. For some people, only actually carrying out formal tests would make the judgments “objective”. To say that subjective judgments can be soundly based and consistently trustworthy is to imply that they would be similar to those made by comparably qualified and experienced practitioners. Professionals who consistently arrive at sound judgments are effectively “calibrated” against their competent peers and also, in professional contexts, against any relevant socially constructed external norms. However, subjective judgments by individuals can also be poorly based, erratic and unreliable.

Another twist is that in some circumstances quite different judgments may be equally appropriate for different purposes.

In many educational contexts, including music and other creative and performing arts, the quality of a work or performance simply cannot be subjected to measurement, standardised analyses and tests. The sole appeal is to complex human judgments. Appraisals are treated as trustworthy when it is known (or has been verified) that several expert appraisers in the field making judgments about the same diverse collection of works or performances show relatively small variations in their decisions. It would surely then be reasonable to trust solo appraisals by one of those experts because their judgments would be known to have a validity extending beyond that of the individual judge. Another way of expressing that is to say that each judge's personal "decision space" has significant overlap with the personal decision spaces of other competent judges, even though the boundaries might differ slightly. When subjective judgments are treated with suspicion because they are inherently unreliable, it is often said that research proves that is so. However, a significant proportion of the research studies have treated the judges as if they were already well tuned to each other even when no attempt has been made to find out if that is the case or to bring consensus about through training, discussion and collaboration on judgments. In academic environments, a full range of attitudes and conventions exist in relation to consensus moderation processes. At one extreme, each academic is regarded as expert in their field, and on that ground alone is competent to make judgments about the quality of student works without further ado. Highly performing or highly qualified individuals have the "right" accorded to them to judge as they see fit. At the other extreme, professionals in many fields as part of their normal practice are constantly consulting about their judgments, and are more likely to be tuned to one another's values, aesthetic sensibilities and norms.

The debate between holistic versus criteria-based judgments (the latter have also been called "analytic" since the early 1900s) has similarly been a long-term issue, at least since Socrates tried to insist on being told the characteristics of piety, so that these characteristics could be converted into rules to determine whether or not a particular person was pious (Dreyfus & Dreyfus, 1984). Applying a fixed set of criteria is no doubt systematic, but to equate the resulting appraisal to being objective is a mistake. Subjective judgments are denigrated in many quarters as being basically matters of taste or preference, whereas the use of formal criteria has the appearance of being more objective, and perhaps more scientific. It is true that if a batch of essays is given to a group of markers who each mark each essay, disagreements are common, even radical disagreements. The results of such trials are then often generalised to attack subjective judgments in general. This then leads to the "conclusion" that standardising the criteria and then training appraisers in the meaning and use of the criteria leads to objectivity. However, as indicated in the preceding paragraph, many of the trials carried out have made no prior attempt to "tune" the judges to a common understanding of what high and low quality works are, and the grounds for appraising them as such. An underlying assumption seems to be that if a large-scale complex judgment is "decomposed" into a battery of

smaller-scale judgments, the judgments become more robust and reliable across judges. What in fact is happening is that a large-scale subjective judgment is being replaced by smaller judgments *which are also subjective*. Does combining these automatically make a less subjective judgment? If so, what is the logic behind this, and how could it be demonstrated in practice?

Holistic appraisal, on the other hand, leaves the set of criteria open. When a musical performance is being appraised, it is listened to as a whole. Particular aspects may be noticed as especially finely nuanced and performed whereas other aspects may jar and detract from the performance. Aspects that are routinely expected and present do not call for special noting or comment—they are implicitly regarded as within normal bounds, tolerances or variations and are, and safely can be, taken for granted. Only aspects that affect the quality of performance call attention to themselves. What have just been referred to as “aspects” match the appraisal criteria that are evoked—both are related to observed properties and in effect offer slightly different angles on the same process. The more expert the listener-critic is, the better their powers of detection and discrimination, which is to say, the greater their sensitivity to the cues that matter. This type of high-order appraisal capability does not come without a great deal of practice. Practice requires exposure to different levels of overall quality, and exposure to equivalently valued variations. A further characteristic of this sort of holistic appraisal is that the features or aspects can be meaningfully discussed, and discussion is possible only when the same terms are attached to the same cues. Learning the vocabulary of appraisal descriptions and explanations is a key necessity. In practical music-learning settings, so much of what has just been described is normal and accepted; openness to criteria, and appraisals using non-identical sets of criteria for different interpretations of the same musical composition are taken for granted.

That thinking needs to find its way into judgments of other types of works as well. Whenever a rubric is made up, a set of criteria is settled upon. In practice, this set always represents only a small sample of criteria out of all the possible criteria. This larger pool may be up to ten times greater than the number selected (Sadler, 1989). Most of the possible criteria are never collected together in one place, but they are there. To choose 10 even from a pool that is, for the sake of argument, only 20 in size makes it fairly likely that some student works of the same generic type and being evaluated at the same time for the same purpose will “need” to draw from the other 10 to properly assess their quality. Which ones they turn out to be cannot be decided in advance; they can be discovered only when appraisals are actually under way. The situation is quite different with mass-manufactured objects where strict conformity to the design is the mark of good quality control. Education in general, and music and the creative and performing arts in particular, are not like that. This is a major flaw in the policy of mandating fixed sets of criteria. Other flaws exist as well, but for these the reader is referred to Sadler (2009).

Holistic appraisals allow the person making the judgment to observe the whole performance through eyes that have a strong “appreciation” element to them. They can see or sense how the whole work is coming (if it is a live performance) or has come (if it is an artefact of some kind) together. If it fails to come together as

whole, or alternatively comes together brilliantly, astute judges know why and the proper way to remark on that. Aspects that are instrumental in reaching a decision about quality are noticed, experienced or felt. Most may well be identifiable, have labels, and be communicated in words to others who understand the context. In some cases, it may be impossible to say exactly why a particular judgment is made or a conclusion reached. Does that make it purely a figment of the judge's imagination? Competent judges tuned in to the context and tuned together on judgments simply may not be able to communicate the full extent of their appraisal knowledge, yet they may agree consistently. In that case, it can be deduced that the judges concerned evidently possess some subtle "knowledge" which they are not able to articulate. Sadler's (1980) exploration of the issue of the domain of a person's fullest knowledge and the domain bounded by their powers of articulation are not necessarily coextensive. This means that they do not necessarily have the same boundaries or map exactly on to each another. Sharing of capability in judgments is improved when judges share not only similar linguistic frameworks but also similar frameworks of experience.

What does all this have to say about assessment in higher education? It is this. It is unfair to give students the impression that a set of criteria, specified in advance, will do justice to appraisals of all student works which are responses to the same assessment task. Apart from the issue of justice, the practice also encourages students to focus on scoring well on all the nominated criteria, so that they score well overall. However, it can also inhibit their ability to see and appreciate their own work *as a whole*. That in turn slows down the rate at which they can learn to engage in continuous monitoring of the quality of their own works during production.

If experienced judges in any field can come to a point where their subjective judgments about quality and the reasons for those judgments are in broad agreement, they exhibit what is known as "intersubjectivity". This has two implications. First, there is no determination of quality, value or worth that can be based on anything more fundamental than that. If someone asks for proof that holistic judgments are fuller and more authentic than can be achieved with sets of criteria, or the reverse, the immediate question has to be: What is the true or accurate appraisal with which the result of applying the two approaches can be compared? In many fields, the definitive judgment is and can only be the collectively tuned subjective judgments of experts. There is no benchmark more fundamental than that. That might tempt some in music and the arts to simply proceed as normal, but instead, it should be seen as highlighting a clear obligation to engage in the tuning process among academics themselves, and to figure out how to induct students methodically and deliberately (as producers and budding practitioners) into the same capabilities.

2.4 Feedback

This is the second front where a reversal is required. Feedback has two broad functions following a judgment about the quality of a student piece or performance. The first part of the first function is to represent the judgment in a simple compact way. This usually consists of creating a symbolic code such as a rating, mark or grade. The second part of this function is explaining how that decision was arrived at, especially pointing out strengths and weaknesses that played a part. The second function is constructive—advice or suggestions as to how the response or performance could have been improved—with an eye to helping the student improve future works in a similar vein. Comments about the current work are likely to be specific, but those specifics can then be used as a springboard for more general advice identifying broader principles. This second function is part of what Bjorkman (1972) called *feedforward*. A scan of the advice published on how teachers should construct good feedback to students shows some consistent patterns—informing students about: how their work rated as a whole; the strengths and (in a non-emotive way) deficiencies; what would have improved their work; and what to do next time. Throughout, it is important to remember that “feedback is an evaluation, not of the person but of the performance in context” (Nicol & Macfarlane-Dick, 2006). Putting the actual content of feedback messages to one side, what is the dominant communication process in feedback? It is all about *one-way telling* (Sadler, 2010b, 2013). That is precisely its greatest limitation; learners do not always learn much purely from being told, even when they are told repeatedly in the kindest possible way.

Written feedforward and feedback have the common characteristic that they rely on messages from the teacher to the student, yet many students, for whatever reasons, seem not to make as much use of them as one would hope. In some contexts, particularly with written term papers and assignments, each student gets to see only their own work and the feedback on it, and possibly that of a few friends. In musical performance and other fields of the creative and performing arts, what students produce is on display for all to see. Criticisms and suggestions flow freely, and formal written feedback is less common. Learning from assessment events requires more than receiving feedback or observing other students’ productions. The facilitative value of peer assessment (Boud, Cohen, & Sampson, 2001) and of enhanced communication through dialogue and conversations (Nicol, 2010; Nicol, Thomson, & Breslin, 2014; McConlogue, 2012) have been widely canvassed in recent years. However, mere engagement with recommended processes is not enough. Those processes have to be geared to a clear purpose. An example of a clear purpose driving processes is described in Sadler (2013).

Another way of analysing feedback is by looking at the main actors and their roles. Consider an air conditioning system. Its control mechanism has a number of components with specialised functions. In order, these are: a device which can be set to and hold the desired temperature, the set point; a thermometer which continuously senses the air temperature; a comparator which measures the

difference between the air and set point temperatures; a mechanism which is triggered when the deviation exceeds a certain amount; an electrical circuit which sends a special code to an external machine (the compressor-evaporator) which heats or cools air whenever required and feeds it into the rooms. Most of the components except for the last one are contained in a little box on the wall. Educators' feedback is similar in some key respects to this air-conditioning control system. In many contexts, the teacher takes the box-on-the-wall role during the learning process. The teacher knows the set point (what high quality consists of), acts sequentially as the sensor, the comparator, the signal generator (feedback writer) and designer of the change that would be needed to bring the student work to an improved state. For a learner to become a self-monitoring producer, most of the elements of their "control system" have to be translated from the teacher's brain as the box-on-the-wall to the students' own brains. The same functions have to be taken over, developed and mastered by the student, not left with the teacher. Unless students are immersed in contexts which facilitate the transfer, they will remain dependent on the teacher.

2.5 A Way Forward

What does this all have to do with the theme of "backwards explanations"? Just this. Suppose the teacher-assessor makes a holistic judgment, and then proceeds to explain it. During the latter process, the assessor invariably makes use of criteria because criteria, by definition, form constitutive elements of evaluative discourses. They work with an open set of criteria and routinely invoke whichever criteria are salient to a particular judgment. This means that they are sensitive to which criteria are relevant and which are not. They know what to do and how to act through extensive experience, along with discussions involving other assessors. This is the appraisal capability that students must develop, and they can learn to do that through the same processes—actual practice and discussion. If students make holistic appraisals without set criteria, represent their judgment in some neutral way (such as 'X' positioned on a "quality scale" with no numbers attached), write down the explanation for their judgment, and provide constructive advice to the producer as to how the work could be improved, and only then engage with fellow students and teachers or markers, the two reversals forming the theme of this chapter can be worked on as a single process. The reason for the definite steps of requiring student appraisers to commit themselves to specific judgments, analyses and advice before discussion is that this reduces the likelihood of groupthink. The combination of first teasing out and explicitly articulating the reasons for judgment and only then engaging with others generally provides significant learning. The goal is not just to learn that differences in judgments occur, but to debate the judgments and the reasons with the aim of arriving at consensus. This discussion needs to include an expert judgment as well. If the circumstances are favourable, and students keep an eye to the need for their judgments to be broadly consistent

with those of expert assessors, the students start on the path to more informed self-monitoring and, if they go the full distance, to the beginnings of connoisseurship.

Where and how can students arrange to get the extensive experience required? It has already been mentioned that in some creative and performing arts contexts, other student works and performances are constantly before other students, and it remains only to turn this into appraisal opportunities. In other contexts, exposure to other works or performances is not readily available without special organization, an instance of that being one-on-one instrument tuition. The ideal arrangement would be to provide students with substantial evaluative experience not as an extra but as a strategic part of the teaching design. The overall aim would be to induct students into knowledge of both the explicit and tacit kinds that would enable them to recognize or judge quality when they see it, and defend their judgments. Additional benefits would likely include deeper student engagement with the content and structure of the academic work and practical performance; a downplaying of teacher-constructed feedback as the critical element for improved learning; and development of the knowledge and skills for high-quality appraisals of the types that are valued in advanced studies or careers after graduation.

References

- Bjorkman, M. (1972). Feedforward and feedback as determiners of knowledge and policy: Notes on a neglected issue. *Scandinavian Journal of Psychology*, 13, 152–158.
- Boud, D., Cohen, R., & Sampson, J. (Eds.). (2001). *Peer learning in higher education: Learning from and with each other*. Sterling, VA: Stylus Publishing.
- Dreyfus, H. L., & Dreyfus, S. E. (1984). From Socrates to expert systems: The limits of calculative rationality. *Technology in Society*, 6, 217–233.
- McConlogue, T. (2012). But is it fair? Developing students' understanding of grading complex written work through peer assessment. *Assessment & Evaluation in Higher Education*, 37(1), 113–123.
- Nicol, D. (2010). From monologue to dialogue: Improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, 35, 501–517.
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218.
- Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment & Evaluation in Higher Education*, 39, 102–122. doi:[10.1080/02602938.2013.795518](https://doi.org/10.1080/02602938.2013.795518).
- Sadler, D. R. (1980). Conveying the findings of evaluative inquiry. *Educational Evaluation and Policy Analysis*, 2(2), 53–57.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119–144.
- Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading in higher education. *Assessment & Evaluation in Higher Education*, 34, 159–179.
- Sadler, D. R. (2010a). Fidelity as a precondition for integrity in grading academic achievement. *Assessment & Evaluation in Higher Education*, 35, 727–743.
- Sadler, D. R. (2010b). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, 35, 535–550.

- Sadler, D. R. (2013). Opening up feedback: Teaching learners to see. In S. Merry, M. Price, D. Carless, & M. Taras (Eds.), *Reconceptualising feedback in higher education: Developing dialogue with students* (pp. 54–63). London: Routledge.
- Scriven, M. (1972). Objectivity and subjectivity in educational research. In L. G. Thomas (Ed.), *Philosophical redirection of educational research* (71st NSSE yearbook, pp. 94–142). Chicago: National Society for the Study of Education.

Assessment in Music Education: from Policy to Practice

Lebler, D.; Carey, G.; Harrison, S.D. (Eds.)

2015, XV, 296 p. 36 illus., 16 illus. in color., Hardcover

ISBN: 978-3-319-10273-3