A Momentary Review of Assessment Principles

Stephen W. Draper
Department of Psychology
University of Glasgow
http://www.psy.gla.ac.uk/~steve/

INTRODUCTION

This paper is a personal review, valid only for today, of what I understand about assessment principles. (As with the rest of the conference, I'm only thinking about Higher Education, and I'm using "assessment" in the British sense to refer to judgements of learning outcomes i.e. roughly of what learners do, not evaluation of what teachers do.) It is written in May 2007 for the REAP online conference. At this point we have had two years to reflect on and improve the principles about assessment that were used in the project proposal. Nearly all the data has been collected, much has not been analysed, and almost none has been reflected on and digested. I feel I know some things that I didn't understand at the start of the project, I know some things I must work out but haven't yet, and I assume my understanding will be significantly further forward in, say, another year's time. This conference is a chance to seek feedback, contradictions, and other challenges to my evolving understanding: so here is a snapshot review of my own ongoing developmental process.

David Nicol and I chose to write our papers without reading each others' drafts. This will make our discussion during the conference important for us personally by satisfying the condition of different starting beliefs that the literature tells us is necessary for us to learn and develop during it, and it will help launch wider discussion by others that is the core business of the conference, and from which its chief value to us will come. I can expect our views to be different because our experience within the project is not identical. The main influences that have driven and are driving constant revision of my best version of assessment principles are: a) constantly seeking to integrate everything I know about education: does anything show that a given version of the principles is wrong or incomplete?; b) talking to students informally or as part of my teaching or from evaluation activities such as focus groups and questionnaires; c) the experience of designing a generic questionnaire "AFEQ" (Assessment and Feedback Experience Questionnaire) based on a version of the principles; d) looking at some other sets of principles for assessment, which I have gathered for reference on a small set of web pages at:
http://www.psy.gla.ac.uk/~steve/rap/principles.html

What makes a good principle? A preliminary answer is a principle that is practicable to implement, but is not yet common practice (so it's worth saying), and can be applied across contexts and particularly across disciplines.

A TALE OF 3 LOOPS: WHAT IS IT THAT STUDENTS NEED FEEDBACK FOR?

The original REAP project proposal was based around the 7 Nicol principles (http://www.psy.gla.ac.uk/~steve/rap/nicol7.html). These could be summarised as all

fleshing out a single aim or super-principle, which might be expressed as "Ensure the supply of information that learners can and do use to improve their knowledge and skills".

Relatively soon after the project began, it became clear that this entirely missed the issue of what Chickering & Gamson (http://www.psy.gla.ac.uk/~steve/best/chickering.html) call "time on task". David Nicol examined a set of 11 assessment principles by Graham Gibbs, and selected 4 that expressed this issue: what I shall call Gibbs' 4 principles (http://www.psy.gla.ac.uk/~steve/rap/gibbs.html). These principles address a different issue: not technical information that a pre-motivated learner can use to tune and correct their knowledge, but how to get students to engage in regular learning work. The overall aim or super-principle might be expressed as "Organise a programme of steady work productive of learning". What students are using feedback for under the Nicol principles is to adjust (regulate) the content of their learning, while under the Gibbs principles it is time and effort that is being regulated. This relates directly to the issue of time management that in first year is one of the common problems that drive students in trouble to seek help from advisors. It also relates to the literature showing that many students ignore expensive written formative feedback and pay attention only to their marks: formative feedback only helps students adjust content, but they need marks (or in general, measures of success) to allow them to regulate their effort.

It is revealing that we didn't initially grasp this. The whole HE industry is organised around disciplines. One consequence of this is that academics' natural assumption about students is that they are dedicated to their discipline, and hence that in learning, the only issue is how to get better at that discipline. In fact students are required to timeshare between multiple tasks, and often between subjects as well. Even if they worked 24 hours a day, and devoted their effort entirely to learning and to no other aspect of life, they would still have to choose how much time to spend on each task. They need feedback on how good their decisions on this are, and assessment should be designed to give them this information.

However there is a third major kind of self-regulation our universities require students to do: choosing subjects. I first began to grasp this when during a discussion with a colleague over the design of the feedback sheets our department used, we casually asked a student what she thought of the feedback on those sheets she had got in the first year. She hesitated, and then said "Well, all I wanted to know was whether I was a psychology student, or whether I should switch to Geography". (Of course our feedback hadn't helped in any way with this).

Academics usually forget this: we have made our choice, and prefer to assume students are apprentices who have also made their choice. However all universities in fact allow course changes, and many are structured in order to require students to take numerous subjects at first and then make a choice. The university sees this as a feature, but academics generally do not support students in this central learning function (despite the research showing that mismatches between students and courses is a major cause of dropout). This is another piece of self-regulation, and assessment for learning implies that assessment should be designed to support this choice. REAP so far has nothing to say about this major aspect of assessment in HE. It is a third feedback loop. It is also notable that Ecclestone (http://www.psy.gla.ac.uk/~steve/rap/eccle.html), addressing a Further Education audience, naturally does have assessment principles focussed on this need. The corresponding overall aim or super-principle might be "Provide students with the guidance and information about both their own abilities and the available courses to support successful decision making by them".

This section expresses a worry based on the cybernetic metaphor for feedback: that there is not one but a number of feedback loops (only the most obvious three of which were

discussed above), each supporting the regulation of a different factor. This line of thinking seemed to imply that we might need a separate set of principles for each such loop or factor. While it is important to check that all are being addressed, pilot work on AFEQ indicated that in practice students had no trouble in general drawing information about multiple factors from a single assessment activity, so principle proliferation is perhaps not the worry it seemed.

RESTRUCTURING FOR AFEQ

Work with Pippa Markham on designing a questionnaire (AFEQ) intended to express the 11 Nicol and Gibbs principles led to an alternative way of viewing the issues that can be helpful. A large part (but not all) of the principles can be seen as the product of two sets: a set of alternative sources of feedback, and a set of steps or elements within a feedback loop.

Traditionally, academics assume "feedback" is what teaching staff provide. An important part of the thrust in the Nicol principles however is to promote self-assessment and peer-assessment. This is for several distinct, but mutually reinforcing, reasons. One is that lifelong learning will require a much greater degree of autonomy and less dependence on teachers. Another is that the great expansion of higher education without the possibility of a proportionate increase in the number of HE teaching staff means that other sources of individualised feedback must be used. A third is that sometimes these other sources are actually better: working it out for yourself is probably better (when it works) than depending on teachers; a peer's explanation can often be better adapted to your understanding than a staff member's.

This suggests three sources of assessment and feedback: self, peer, teacher. These should all be considered as sources in each case and context.

A fuller list of sources would be:

- Self as in working out what is required, whether a given answer is satisfactory, etc.: i.e. dynamic analysis or "reflection".

- Self as in bringing past knowledge from previous courses to bear on this one. This is particularly important in bringing existing understandings of assessment criteria e.g. what defines a good essay in this discipline, whether and how you should give error bounds on numerical answers, etc. Many courses (quite rightly) rely heavily on prior unstated understandings by students of required standards.

- Peers.

- Teachers as expressed in written, stable documentation. Students value, for example, written course documentation specifying assessment criteria, or example essays with comments.

- Teachers expressing information informally, orally. Students also go to classes partly to hear informal descriptions of what is required, what will be in the exam etc., and to tutorials to get more personalised comments.

- External sources such as the published literature in the discipline. In courses where students are required to search out and read the literature this also becomes in part a model of what is required for both content and criteria. It is a telling question to ask of any course whether the staff really only want students to reproduce their opinions, or to relate to a wider academic community in the subject. Also as more online tests become available, students may increasingly do tests set by others outside the course.

The elements in a learning feedback loop may similarly be enumerated, and then each particular course design may be interrogated for how well it supplies each element. The most obvious element is a judgement by the teacher (or peer, or ...) indicating a difference between what was done and what was wanted. However as Sadler has emphasised, this cannot be of any use to a student who does not understand what the desired standards (the assessment criteria) were. It's no good telling a student their essay should "be more critical" if they do not understand what being critical means (in this discipline). Thus understanding the criteria is another necessary element in the loop. Another basic point is that unless there is an opportunity to apply the feedback to doing the task better, nothing can be gained. Finally, it can be worthwhile to apply the medical analogy to feedback and distinguish identifying symptoms, diagnosing the underlying disease, and prescribing a remedy. Putting this together suggests these generic elements to a learning feedback loop:

- Understanding the task/assessment criteria (what is to be aimed for)

- Support and advice during the task: if students get stuck, they will ask questions and need answers before they have produced anything. Project supervision is a familiar example, but so too is provision of help desks or tutors in "labs" for doing maths, computer programming, physics problems, etc.

- Generating a judgement e.g. a mark and/or "symptom" reports about what is less than perfect.

- At this point, as Dweck has demonstrated, there is a crucial but overlooked judgement made by the learner. This is a judgement of what the feedback is telling them about: is it random noise, a measure of their ability (i.e. a trait they can't change, but which might indicate they should drop out or change course), information about how to improve their skills, information that about whether they made enough of an effort, ... Dweck also demonstrated that teachers could change this judgement easily by the words in which they expressed their feedback.

- Constructive suggestions about how to change what was done (cf. prescribing remedies). These may be supplied by the person given feedback, or by the learner (after all, if we know we misspelled a word we can find out how to correct it ourselves: we only needed to know what the problem was, not the remedy).

- Opportunities to act on the feedback.

MAJOR OMISSIONS

The 11 principles that REAP has so far focussed on (or their rejigging into the 10 David Nicol currently promotes) omit a number of major issues. These include:

- Support for choosing subjects of study and disciplinary identity ("the third regulation loop" in an earlier section).

- The crucial nature of support during an assessed task, not merely feedback after it's all over. For example many subjects such as maths, computer programming, physics are largely organised around "doing examples" or problems. Doing 5 a week would be unexceptional. Many students can unstick themselves, for instance using the textbook, when they run into problems. However those who don't cannot do any work until they get help. In a typical course, this help is available from staff at most once a week. The weak students can only learn during this one window: obviously they will quickly fall behind and fail if they cannot get all 5 problems done at that time. However the workload is usually designed so that a weak student cannot do it all during the lab or tutorial. In other words, this function of supporting learners during the task is typically mission critical (unlike the lectures, for which reading the textbook is very often an adequate substitute).

- Accreditation, and all the issues around summative exams. These are of the first importance to universities and students alike, although often not to the academic staff.

- Authenticity: whether (assessment) tasks relate to contexts and practices in the outside world of importance to the learner and/or discipline.

- Alignment: whether (assessment) tasks relate to the stated learning aims; and whether they should. Rowntree's principle (http://www.psy.gla.ac.uk/~steve/best/rowntree.html) that we should assess what learners have learned, not just what teachers want them to learn, runs counter to this, as does the belief by some that many of the valuable things gained in higher education never do and never can appear in explicit pre-specified learning aims and objectives.

So how might we assess the adequacy of the REAP 11 principles? Compared to other sets I have seen, they seem good. Compared to an ideal of being comprehensive, the above list of omissions shows they are seriously incomplete. However we might argue that from the viewpoint of making a practical impact, simply repeating principles that are widely attended to (e.g. making summative tests fair) will distract attention from a selection of a few that might make the biggest improvements in today's common practices. On this practical view, the 11 principles' most important messages are probably 1) Make much greater use of peer-assessment and self-assessment practices. 2) Do more to ensure learners understand and can actively use the assessment criteria and concepts.

Besides all these ideas, however, which mainly come from established literature or the restructuring of the ideas in the Nicol principles around more explicit use of the feedback loop concept, there are some others emerging from strikingly successful learning designs. It thus seems that these might score the highest in practical impact, and in having been omitted from REAP and other published principles.

THREE EMERGING PRINCIPLES

In this section I sketch a formulation of three principles that seem to me to be emerging as crucial in explaining the best learning designs we have seen both in the literature, in some cases submitted to this conference, and within the REAP project. In other words, in considering particularly the theme of great designs, we can ask are there any underlying features or principles that might characterise what is important in them. These are the proposed assessment principles I myself would particularly like feedback and comment on.

Contingency & responsiveness

Nicol's principle no.7 is about having assessment provide information to teachers that can be used to help shape their teaching. There is nothing in this about quantity, but I've noticed that a significant feature of many innovative designs of striking effectiveness is that the amount and importance of such learner to teacher feedback is enormously greater (by two orders of magnitude) than in conventional designs. Most courses pay some attention to exam results and assessed exercises, and may adjust teaching the following year to address topics that seem to have given particular poor results. This is a response cycle time of one year, and may result in minor adjustments to content. In contrast, in Just In Time Teaching every class is used to respond to learner test results and questions: a response cycle time of at most a week, perhaps every few days; and in principle the whole class content (not minor aspects) depend on it, which is approaching 100% contingency. In using the Mazur method with EVS (electronic voting systems), the overall lesson plan may be fixed, but the nature and amount of teacher explanations are substantially tailored to the class response to the last question. This is considerable contingency over the short scale, and affecting a

part of every lesson. Some other uses of EVS have used in effect a diagnostic tree structure of questions to home in on what that particular class needs to work on: a higher degree of contingency. In Baxter's case study at this conference, the weekly set tasks are only written by the teacher a few minutes before they are announced, and are adjusted to the state of the class at that time.

In summary, there are two parts to this principle: a) responsiveness or frequency, i.e. how often such adjustments to the teaching are made, with standard practice being once a year, but a number of innovative designs arranging to do it every week; b) contingency or size of the change: how much of the content of teaching is made to depend upon prior information from students such as their responses to quizzes, or submitted questions. Informally it may also be said that students feel much more part of a class when it is visibly responsive to the state of the students. In fact lecturers who use humour or other comments to show awareness of the class feelings and situation seem to gain much higher ratings than others, even though in fact they are not adapting their teaching in a way directly relevant to learning content.

This relates to the abstract ideal of one to one personal teaching, which would be completely responsive and contingent. What is notable about the learning designs mentioned is that they are ways of achieving these properties with very big classes. They depend upon creating a functioning channel from learners to the teacher. If this is from quizzes, it requires creating the questions. If it depends upon volunteered input from students, it requires not just having a channel, but working hard to create the atmosphere in which sufficient students feel it is comfortable and worthwhile to send in comments and questions. This may require significant effort from staff, but it seems to bring significant learning benefits.

Solo:group work relationship

I have finally grasped from my students that productive learning does not come from seeing individual (solo) work as an alternative to group work (as the list of sources in an earlier section might seem to imply), but rather from setting up the right relationship between them. In the end, a learner is judged by everyone (in our culture) by what they can do by themselves, and only learns this by doing the task alone. On the other hand, working in and with a group can provide various benefits much harder to get alone. One of these is discussing concepts between people with initially different views, as this is a strong prompt to producing not just "answers" but reasons and counter-arguments for views. Another is getting expert help to cover a small gap in one's knowledge: a very frequent need for instance in learning computer programming. Another is checking out one's plans for learning and time management, particularly for a new task where you have no previous experience to rely on.

One successful recipe is for each group member to work solo on a set of problems (in say, statistics or maths or physics) until they get stuck; and then bring their remaining problems to the group for "unsticking". The converse pattern, also successful, is that in the Bates case study at this conference: working in a group on the first few of a new type of problem, so that initial problems are dealt with and some confidence accumulated, followed by solo work on the remaining problems to check that each student can now do them without help. In contrast we have seen problems in groups without these recipes: either resenting others exploiting their work and feeling they would do better solo, or else enjoying groups but never working solo and usually doing poorly in exams later since they have never attempted a task unsupported.

Although some students and some course designers have found successful recipes intuitively, many of the rest of us do not find this principle intuitively obvious. There are probably at least three underlying concepts we have to correct to clear the way for it.

Firstly: learning generally involves a progressive withdrawal of scaffolding, as discussed in the next section. It is not enough only to "go through" a task with others helping; not enough to "see one" without then proceeding to "do one". Secondly, most group work in all settings other than academic ones is about division of labour ("collaboration" in some literature) to achieve a jointly produced outcome. The point of division of labour is NOT having to do, or to know how to do, each other's work. Essentially all the group-work literature in business, management, and most psychology concerns only this type of group. However learning groups should have a different aim: to end up with everyone being able to know and do the same things independently ("cooperation" in some literature). Many learning groups are unproductive perhaps because they naturally and unconsciously adopt tactics suitable for non-learning groups. However you cannot learn for someone else; and you may not really be able to read for someone else. You can however do other things for them: test them, offer them alternative explanations, exercise their knowledge (which reading does not do), and so on. Thirdly, I, and I suspect others, have misunderstood Tinto's distinction between social and academic integration. In fact, as has often been reported, students seldom have problems making friends or having a good social life. What is less routinely the case, however, is students having people and occasions on which they discuss academic issues. Thus social activities, ice breakers, freshers' parties etc. are "social" in the everyday sense, but are not addressing the key issue: working together on academic (not social) tasks. A subtle but important correction to interpreting Tinto is to consider social-academic as referring to the solo-group axis, and considering that every learning activity may be done in either way, and that the most productive is probably a judicious mix in which the two modes are brought into a productive relationship. For instance it is not really enough just to spend some time in each mode: what is best is to prepare by solo work for what to do in a group, and then follow up the group work by solo work. The Baxter course design brings this home because the participants never meet in the flesh, only online. Many of them express at the end of the year how good they feel their group was: but it isn't friendship in the usual sense of meeting and socialising together. It is however true "social integration" or "learning community" in the sense of their learning being better and feeling better because peers are involved centrally in it.

The Bates case study illustrated one recipe that is explicitly designed into the course, by requiring both a phase of group work, and then an enforced and assessed phase of solo work on each set of problems. The Baxter course design goes to some effort (following problems observed in the preceding pilot project) in specifying to each group how they might divide up the work, and requiring that each person does a different part solo. It requires them in the larger tasks then to assemble the parts into a single essay. However there seems to be at that point considerable variation between groups, and only the best minority go well beyond simply pasting them together, and have a real group discussion on improving the joint product.

Learner pro-activeness

The REAP proposal and the original presentation of the Nicol principles explicitly mentioned the aim of promoting "learner self-regulation", but did not show the links between this aim and the specific principles. In fact they are independent but fully compatible issues: for each principle, you could promote it with a low or with a high degree of learner autonomy or "proactiveness". Thus, for example, a teacher may promote peer assessment by organising and requiring such a process in class under supervision, or they may (as my department does) strongly recommend that students form their own study groups but do nothing to schedule this, or they may not even mention it although in some circumstances it

will happen anyway: for instance mature students typically seek each other out and work together a lot, and in computer programming, students find it natural and necessary to ask each other for help when they need it. Similarly a student may get teacher feedback without asking for it on some exercises; or they may take advantage of published staff office hours to ask for it; or they may approach the lecturer without invitation at the end of a lecture. This applies more generally to every learning activity whether or not it is assessment: some students will only read what they are told to, others will read widely; some will only do compulsory exercises, others will set themselves such tasks. For every aspect of the learning and teaching process, including all aspects of assessment and feedback, we may apply a scale ranging from whether the teacher takes full responsibility for making it happen and enforces it, through recommending it, to not mentioning it, to actively obstructing it (e.g. ordering students not to read texts outside the recommended ones).

Of course as long as the teachers are organising it, there is no need for learners to do so or even to think about it. Thus if we adopt the aim of promoting learner autonomy and self-regulation, then we should probably adopt a policy of progressive withdrawal of scaffolding: begin by providing them with fully organised activities, and then leave more and more of it to the learners to arrange. The dual facets of learner proactiveness and withdrawal of scaffolding are connected at a profound level. In the classic cases of scaffolding, for instance in an adult helping a child with a jigsaw, there is not in fact any real question that the child can do all the actions (manipulating the jigsaw pieces): what they can't initially do is chain together the actions into a productive sequence. In other words, scaffolding is largely a matter of skill at planning and management rather than of component subskills. Similarly in HE learning, students may not really appreciate the importance of each activity at first (for instance, how and why peer assessment is useful), but can progress from compliance to organising it themselves.

Implementing the progressive withdrawal of scaffolding often requires contingency based on feedback from learners to teachers. Course designs in HE frequently fail to show any variation in the amount of scaffolding: misplaced ideas of "consistency" lead to what is in fact, for learning, worst practice, like keeping a plaster cast on a leg year after year instead of taking it off at the first possible moment to start exercising the muscles. Individual course designers however sometimes show a good practical grasp of it. The Baxter course design has some aspects of this, with set tasks later on in the year coming with less explicit support and breakdown into component tasks than earlier ones.

CONCLUSION

Formulating the "best" set of assessment principles is an ongoing project. I would particularly welcome discussion of the three emerging principles above.