Driving learning via criterion-referenced assessment using Bloom's Taxonomy

Raymond Lister, Faculty of Information Technology, University of Technology, Sydney, Australia
raymond@it.uts.edu.au

Introduction

The goal of every university teacher should be to help each student realize his or her potential. For many teachers, there is a natural tendency to spend more time with the struggling students, but the problem then occurs that the stronger students are not being extended. The problem is especially difficult in the introductory subjects of a general science degree. In such degrees, students drop subjects as they progress through the years, eventually majoring in only one or two of the subjects they studied in their first year. It is difficult to accommodate, in the same classroom, and within the same assessment scheme, both the students who intend to major in a subject and the students who do not intend to pursue the subject any further. There is also the danger that students who attain only the minimum passing grade, and who elect to continue studying in that same area, are not adequately prepared for the subjects that follow.

Traditional norm-referenced approaches to assessment do not easily accommodate students of different abilities and aspirations. In pure norm-referencing schemes, all students in a class attempt the same assessment tasks, and grades are assigned according to some desired grade distribution. Such a scheme does not guarantee minimal competency in the weaker passing students, nor does it guarantee that the best students have been extended.

Both weak/strong students, and also major/non-major students, are better accommodated by criterion-referenced grading schemes. In pure criteria referencing, explicit clear criteria for each grade are communicated to students. A grade is assigned to each student according to the criteria satisfied by the student, irrespective of the resultant grade distribution. Such a scheme allows some students to attain a pass by demonstrating their grasp of the fundamental principles of a subject, while other students have the opportunity to extend themselves by exhibiting greater creativity.

In this paper we describe our use of the criterion-referenced approach to assessment, where the criteria are based upon Bloom's taxonomy. In our Bloom-based assessments scheme, all students in the class must satisfactorily complete a set of assessment tasks designed to demonstrate competence at the Knowledge and Comprehension levels of Bloom's taxonomy. Any student who is content with the minimal passing grade need not complete any more assessment items. Other students, who elect to seek a higher grade, must also complete assessment tasks at the Application and Analysis levels of Bloom's taxonomy. Student's who satisfactorily complete these tasks, and stop at this point, attain one of the two middle passing grades. Finally, students who elect to seek the highest grade must go on to satisfactorily complete further assessment tasks, at the "Synthesis" and "Evaluation" levels of Bloom's taxonomy. Each student is free to decide for him/herself what grade they will try to achieve. They are also free to approach the lecturer for advice, but in practise few do so.

We have applied this assessment system in seven semesters of teaching, five semesters of teaching introductory programming, and two semesters of teaching introductory databases. After we had used this assessment system for a single semester, to teach introductory programming, we published descriptions of this assessment approach (Lister and Leaney 2003a, 2003b). This paper differs from those earlier papers in two ways. First, the earlier papers were aimed at members of the information technology community, and those teaching programming in particular, whereas this paper summarises the broad structure of the assessment scheme in way that is accessible to academics in

non-computing disciplines. Second, this paper benefits from six more semesters of experience with this approach to assessment.

A review of Bloom's taxonomy

Bloom's taxonomy contains six levels, with the organising principle that competence at a higher level of the taxonomy implies a reasonable degree of competence at the lower levels (Bloom and Krathwohl 1956). Furthermore, successive levels of the taxonomy are paired, to form three groups, with qualitatively different assessment standards expected between the different groups. The remainder of this section briefly reviews those three groups, from lowest to highest.

Knowledge and Comprehension

Informal descriptions of the taxonomy frequently confuse the two lowest levels of the taxonomy. At the lowest "knowledge" level, a student can regurgitate a fact when prompted for it, without necessarily understanding the significance of the fact, a level of competence that can simply be achieved via rote learning. We do not advocate the encouragement of rote learning. The next level of the taxonomy is the "comprehension" level. It is a higher level because a student competent at the comprehension level understands the significance of a fact. A student manifests that understanding by supplying knowledge when prompted for it in a way that is different from how the material was first taught. For example, in the case of physics, a student might be required to demonstrate their mastery of a concept by correctly using the concept on a simple problem not previously seen by the student.

Application and Analysis

At these intermediate levels of the taxonomy, students are expected to be able create and analyze artefacts, but within a well defined context. For example, in the case of physics, a student might be required to solve a well defined but non-trivial problem. Bloom (1956) distinguishes between comprehension and application as follows: *A demonstration of "comprehension" shows that the student can use the abstraction when its use is specified. A demonstration of "application" shows that he* [sic] *will use it correctly, given an appropriate situation in which no mode of solution is specified (p. 120).* For example, if a student is asked to solve a certain physics problem by applying a certain equation, then that is an exercise at the "comprehension" level. If, however, it is left to the student to realise that the equation is required to solve the problem, then that is an exercise at the "application" level.

Synthesis and Evaluation

At this highest level of the taxonomy, students are expected to show considerable skill in setting and achieving their own goals, with minimal assistance from the teacher, and also show critical skills in analyzing artefacts. While a student working at the lower analysis level is in a limited sense synthesising something, a student operating at the application level is engaged in a relatively mechanical, algorithmic process. A group of students working at the application level will usually produce fewer qualitatively different solutions than there are students in the group. In contrast, a student operating at the synthesis level demonstrates creative flair. The analogous argument can be made for the difference between the analysis and evaluation levels.

Six Bloom levels into three institutional grades

At our university, there are four passing grades, "Pass", "Credit", "Distinction", and "High Distinction" (or simply "P", "C", "D", "HD" respectively). In developing our grading criteria, we found it relatively easy to assign criteria to the "P" and to the "HD". However, we struggled to devise criteria to distinguish qualitatively between the middle two grades, "C" and "D". (Note the use of "qualititatively" throughout this section.) Eventually, we decided to assign the same qualitative criteria to the middle two grades. Thus, we effectively have three passing grades.

To develop the criteria for each of our three passing grades, we associated each of the three pairs of Bloom's taxonomy with one of the three grades, in ascending order. Thus, a "P" is awarded for performance at the Knowledge and Comprehension levels. A "C" or "D" is awarded for performance at the Application and Analysis levels (whether the actual grade awarded is a "C" or a "D" depends upon the student's work − work that receives a "D" is better than work receiving a "C", but the difference is one of degree, and not a qualitative difference). A "HD" is awarded for performance at the synthesis and evaluation levels.

Table 1 shows the grade distributions for the first time and the most recent time in which the Bloom-based assessment system has been used. The grade distribution for the introductory programming class broadly reflects the grade distributions in the subsequent five programming classes. The high failure rate is not a reflection of our Bloom-based criterion-referenced assessment system, as it is a reflection of the minimum standard we expect of our programming students. The failure rate would have been similar in a traditional norm-referenced approach to assessment.

Table 1. Grade distributions for two semesters in which the Bloom-based assessment system has been used

Subject	Failed	P	C/D	HD
Semester 1, Introductory programming	29%	18%	47%	6%
Semester 7, Introductory databases	<1%	19%	65%	16%

Knowledge/Comprehension and grade "P"

In this section, we describe in detail the types of assessment activities undertaken by all students in the introductory programming class. Consistent with the two lowest levels of Bloom's taxonomy, students are required to demonstrate their understanding of the fundamentals of the subject area, without manifesting significant creativity.

Laboratory exercises followed by a laboratory examination

Our laboratory exercises are intended to require no more than one hour of work (and frequently less). These exercises are designed primarily as learning experiences. However, these laboratory exercises are actually summative, as although they do not contribute to a student's final mark, they must be completed satisfactorily. Our approach to marking is relaxed. We try to maintain the ethos of formative assessment. We merely want to see that a student has made a solid attempt, and learnt something. Students may also make multiple attempts, and seek assistance from teaching staff.

Students may work on the exercises outside the lab, and merely present the exercises at a laboratory session for marking. While the sincere student benefits from this laissez faire arrangement, there is considerable scope for cheating. That was a consideration in also having a laboratory examination, approximately two thirds of the way through semester. Typically, around 10% of the class fail this examination at their first attempt, but almost everybody passes after two attempts. If a student fails on the second attempt, we promise a third attempt at the laboratory examination provided the student passes the multiple-choice examination (see below). In practise, we have never had to run a third laboratory.

Multiple Choice Examination

Most teachers dismiss multiple-choice exams as either being too easy, or as being hard merely by assessing obscure facts. However, if the recommendations of the multiple choice literature are followed (e.g. Ebel and Frisbie 1986; Linn and Gronlund 1995) then legitimate, demanding multiple-choice exams can be set. Furthermore, we set the pass mark for our multiple choice exam at 70%, a

typical pass threshold for "mastery" examinations, since we believe that students must manifest a strong grasp of the fundamentals.

Discussion of pass/fail assessment: the danger of rote learning

When multiple choice exams are used, there is a danger that students will rote learn the answers. This danger is real for poorly constructed multiple choice exams that test students at the Knowledge level, but it is less of a problem for multiple choice exams designed to test at the Comprehension level. To answer question at the comprehension level, the student must genuinely understand the material being assessed. Within educational psychology, a distinction is made between meaningful reception learning (Ausubel 1963) – which we advocate – and rote learning (Lefrancois 1999, pp. 213-219) – which we do not advocate.

In our examination results, there is little empirical evidence of rote learning among passing students. The examination for our second running of the subject contained a mix of questions. A small number of the questions had been seen by students before the exam, and most questions were "unseen" (i.e. were about problems that students had not seen before the examination). The typical performance on seen and unseen questions was comparable. On the majority of unseen questions, 63-85% of passing students correctly answered each of these questions, with a median of 78%. On previously seen questions, 63-89% of passing students correctly answered each of these questions, with a median of 81%. These comparable percentages suggest that passing students do not rely on rote learning.

Discussion of pass/fail assessment: the danger of streaming

A potential concern with the pass assessment is that students awarded a "P" have been "streamed", and are not suitably prepared to take any subsequent subject for which this subject is a prerequisite. The intention of the scheme is not to stream students. It is the intention that students who achieve a 'P' should be able to progress to the next subject. An important feature of this assessment scheme is that the 'P' students are not taught a subset of the syllabus taught to all other students. It is just that the 'P' students are not required (in this particular subject), to manifest their learning of those concepts in as sophisticated a form as the students who achieve a higher grade. Furthermore, as discussed above, the "P" assessment is designed to not reward rote learning, but instead reward genuine comprehension of the concepts taught.

In the introductory programming subject in which this assessment scheme has been used, there is empirical evidence that 'P' students have not been streamed. From the students enrolled in the first class assessed via this criterion-referenced approach, fifty students who gained only a "P" went on to attempt and pass a subsequent programming subject, which used conventional norm-referenced assessment. Of these 50 students, 32% achieved another pass, 61% achieved a credit or distinction, and 6% (i.e. 3 students) actually achieved the highest available grade, the high distinction. This good subsequent performance by "P" students is possible because the criterion-referenced approach aims to provide "P" students with a solid grounding, and genuine understanding, in basic concepts for "P" students. The criterion-referenced approach does not leave "P" students to uselessly flounder, on assessment tasks aimed at higher achieving students.

Application/Analysis and grade the "C" and "D" grades

We assess for the credit/distinction in a single, substantial assignment, which is typical of the type of assignments given in traditional norm-referenced assessment regimes. It is our contention that, in norm-referenced assessment tasks are implicitly aimed at the middle-achieving student, so traditional assessment tasks are suitable for assessing the "C" and "D" grades in our criterion-referenced approach. In our criterion-referenced class, since the "P" grade is entirely determined by assessment items other than this assignment, participation in the assignment does not directly affect a student's

likelihood of failure. Among those who did attempt the assignment, we found the atmosphere less fraught than in traditional classes, where the assignment does influence pass/fail. This calmer atmosphere is more conducive to learning and less conducive to plagiarism.

Synthesis/Evaluation and grade "HD"

Bloom has defined synthesis as "Skill in writing, using an excellent organisation of ideas and statements" (p.169). He further states that synthesis is the "ability to write creatively a story, essay, or verse for pleasure, or for the entertainment or information of others" (p. 169).

Synthesis: Individual Project
For a "HD", students in the introductory programming subject are first required to write a program of their own choice. They are given some simple criteria to ensure that their effort is not trivial. For example, it is specified that the program should contain at least 200 lines of code, and that certain programming techniques must be used. Students in this "HD Community" must also give a short presentation on their project to the other students who are doing projects.

The idea of a self-selected project should be easy to adapt to any science discipline. The project could be tied to the use of certain measurement instruments available in the class laboratories, for which students have already received instruction while pursuing the assessment requirements of the lower grades.

Evaluation: Student Peer Review
The project described above is merely the first half of the HD assessment. After completing their own projects, students must then peer-review and write a report on the project of two other students in this "HD Community".

One very pragmatic benefit of this peer review approach is that it reduces the marking burden on the teacher, but peer assessment has other and more important benefits. One of the benefits is social. The more highly motivated students get to meet each and other, on their own terms. They form a smaller community within the larger class. This "class within a class" is also good for the mental health of the teacher. For the last 2 or 3 weeks of semester, the teacher is working with only the highest achieving students in the class. It is a refreshing change from the rest of the semester, where a teacher in a large introductory science class will probably spend most of their time communicating with the students who are struggling with the subject.

Student Feedback

In the second semester that we ran this criterion-referenced assessment scheme, the class was surveyed as a routine part of the faculty's quality assurance process on all subjects. The students were asked the faculty's standard set of questions, to which they answered, on the common 5-point Likert scale (strongly disagree, disagree, neutral, agree, strongly agree).

One of the survey items was "My learning experience in this subject were interesting and thought provoking". In response 11% of the class strongly agreed, 58% agreed (total 69%), 18% were neutral, and 13% disagreed. One of concerns had been that the 37% of the class who eventually achieved a "P" may have been bored by an assessment diet of lab exercises, lab exam, and multiple choice examination, but this does not appear to be the case.

Perhaps the survey item of greatest relevance to the evaluation of the assessment scheme was "I found the assessment fair and reasonable". In response 7% of the class strongly agreed, 57% agreed (total 64%), 30% were neutral, and 7% disagreed.

Conclusion

A criterion-referenced approach to assessment need not use Bloom's taxonomy, nor does the application of Bloom's taxonomy to assessment imply that the assessment is criterion-referenced. However, this paper demonstrates how the two ideas can be combined powerfully, in a criterion-referenced approach where the criteria are derived from Bloom's taxonomy.

In principal, standard two- and three-hour examinations can be constructed to achieve most of the same goals achieved in our assessment system. Most lecturers aim to set an exam with a mix of questions, testing students across a range of abilities. In practise, however, this aim is often not achieved. The writing or marking of such exams is a lot of work, particularly for the more challenging questions. Consequently, in practise, many exams only test students at the knowledge and comprehension levels of Bloom's taxonomy. Well constructed written exams can test students at the application and analysis levels, but testing students at the more creative synthesis and evaluations levels usually requires longer, less well defined tasks, such as assignments or projects. Even when academics succeed in setting exams that test students at multiple levels, it is not clear what students gain from scoring partial marks on the more difficult questions. The danger is that both teachers and students enter into an unspoken conspiracy, where students commit to paper pathetic answers to difficult questions, and academics contrive to find ways of awarding marks to such answers.

One often hears university lecturers grumble that many students seek merely to pass, and have no desire to achieve any higher. In this assessment scheme described in this paper, where students choose explicitly the grade for which they will strive, the grade distributions given in Table 1 indicate that many students do want to achieve more than just a bare pass. In the introductory programming subject, a majority of students achieved a grade higher than a pass. In the introductory database subject, approximately 80% of students elected to strive for a grade higher than a pass. The factors influencing students as they make these choices is currently the object of research.

References

Ausubel, D.P. (1963) *The Psychology of Meaningful Verbal Learning.* London: Grune & Stratton.

Bloom, B.S. and Krathwohl, D.R. (1956) *Taxonomy of Educational Objectives: Handbook I: Cognitive Domain*, Longmans, Green and Company.

Ebel, R. and Frisbie, D. (1986) *Essentials of Educational Measurement*. Prentice Hall, Englewood Cliffs, NJ.

Lefrancois, G. (1999) *Psychology for Teaching*, Thomson Learning. 10th edition.

Linn, R. and Gronlund, N. (1995) *Measurement and Assessment in Teaching*. Prentice Hall, Upper Saddle River, NJ.

Lister, R. and Leaney, J. (2003a) *Introductory Programming, Criterion Referencing, and Bloom*, 34th Technical Symposium on Computer Science Education (SIGCSE 2003), Reno, Nevada USA, February 19-23, 2003, pp 143-147.

Lister, R. and Leaney, J. (2003b) *First Year Programming: Let All the Flowers Bloom*, Fifth Australasian Computing Education Conference (ACE2003). Adelaide. February 4-7, 221-230.