Chapter 38

Student Assessment In A Problem-Based Curriculum

Dirk Tempelaar

University of Limburg, Maastricht, the Netherlands

Introduction

In this paper two different student assessment systems will be described. Both systems are developed for use within a problem-based curriculum. The first system, the ‘classical’ problem-based learning assessment system, is the system that our faculty of economics and business administration started with, when in 1984 the first students entered our school. This system is called classical because of the fact that it was taken over from the medical school, being inextricably entwined with all the other ideas, experiences and solutions on problem-based learning. However, as with many ideas that are taken over, the system inevitably began to change, from year to year, with continuous shifts in some years, and discontinuous large jumps in others. One of these paradigm shifts occurred in the academic year 1991/1992, the year of the general reprogramming of the faculty’s curriculum (see also the contribution of Hans Kasper, this volume). In that year the backbone of the classical assessment system, the progress test, was removed from the first year program and replaced by a new type of test, the so-called overall test. To stress the importance of this paradigm shift, I will call the assessment system that was evolved by that time and which is largely the regime prevailing at this moment, with the new overall test as backbone, the ‘alternative’ problem-based learning assessment system. In this paper I will not only describe both systems, but also give the main arguments that led to the abolishment of the progress test, which sometimes is seen as one of the foundations of problem-based learning.

Requirements For An Assessment System

A key prerequisite of any assessment system, whether problem-based or not, is that its procedures are congruent with the educational and instructional principles. Students, and especially those of economic faculties, who are daily trained in optimising behaviour, adapt their study approach to the assessment procedures to maximise their chance of success. When designing an assessment system, this fact cannot be ignored. A badly designed system gives students many stimuli that counteract the educational principle (e.g. our tests tempt them to just memorise facts, whereas we want them to apply their knowledge). At the other extreme, in a well designed assessment system we have an additional instrument to achieve the desired behaviour: the test system serves as an educational tool. This line of reasoning makes it clear that we cannot expect a traditional assessment system to work well within a problem based
curriculum. Such a curriculum has its own educational and instructional principles, and the assessment system has to be subordinate to them (see also the contribution of Mien Sengers, this volume, on this dependence of assessment on learning). In the following two paragraphs, I will outline two different assessment systems. Although being different, both systems are designed for a problem-based curriculum; they only differ in terms of the aspects of problem-based learning that are labelled as being of primary importance in the design of the assessment system.

**Progress Test**

A central element in the 'classical' problem-based learning assessment system is the so-called progress test. This is a test exclusively designed for problem-based learning: you won't easily find a comparable test in any other educational system. The reason that the architects of problem-based learning in the medical school spent so much energy in designing a new type of test exclusively for problem-based learning, is that they considered all other assessment systems to be incongruent with the main principles of problem-based learning. These principles are (see van der Vleuten & Wijnen, 1990, my prime source of information for this whole paragraph):

- the principle of self-directed learning, the student being fully responsible for her or his study actions, and
- the learning through practice principle, together with an emphasis on abilities beyond knowledge.
- Knowledge that students can apply and use, e.g. to solve problems. Thus, there is an integration of theory and practice and, along with that, an integration of the disciplines or subject matters.

On the basis of these principles, the medical school made the following choices with regard to the design of its assessment system:
- to centralise all assessment activities in order to facilitate the pursued integration of disciplines;
- to design a comprehensive system, accounting for the need to go beyond mere knowledge, and include competencies in the assessment. The system should regard assessment as an ongoing occurring activity, on a continuous basis; dependence upon momentary information does not fit in the picture of self-directed learning.
- Assessment should serve both summative and formative purposes: it should play a 'mirror' function for students and teachers. In the system there should be a separation of the teacher's role and the role of the examiner, since both roles are incompatible. Finally, what we all aim at, the assessment instruments should be reliable and valid.

Different instruments were used to evaluate the four relevant types of clinical competencies: knowledge, skills, problem-solving, and attitudes. In this paper I will only focus on the first one: knowledge. The other competencies are rather domain specific and play only a minor role in the grading system, even in the medical school.
To assess the knowledge of the students, the medical school designed a dual system. One component of that system is the block test. This is a very classical component: at the end of each block, students get a test consisting of 150 or more objectively scorable items representing the cognitive objectives of the last six weeks of education. Except for the fact that the item format is true/false, and not multiple choice, the format of the test seems rather standard. However, the function of the test is not standard at all. Block test scores do not play any role in the grading of the students. In our terminology: the test does not serve as a summative purpose, only a formative one. The main argument behind this one-sided use of this test is easy to reproduce. Problem-based learning stimulates self-directed learning. This, however, would be a hollow phrase if all blocks are tied up by tests that force students to master the complete spectrum of subjects. Such tests would stimulate a concentration on the subjects that are deemed important by the lecturers who wrote the block book, but that can be far apart from the learning path you prefer. They also stimulate learning by rote strategies and a focus on details, which isn’t quite problem-based learning. The only way to completely circumvent these unwanted side effects of assessment is to exclude this type of test from the grading system.

To enhance the formative value of the block test, students get much more feedback than only their general test score. Using the table of specifications of the block, which explicitly states the objectives for different disciplines, all test items are clustered and the students’ feedback consists not only of the overall score, but also of a more detailed insight in her or his knowledge on all the different subjects of the block. This insight is both absolute in nature (what do I know in comparison to what I ought to know?) and relative in nature (do I know more or less than my fellow students on this subject?).

A second measure to overcome the negative side effects of the block test is the inclusion of the question mark option within the true/false format of the items. This options allows the students to ‘pass’, thereby indicating that they don’t master the subject of the item, probably because they spent their time on other subjects. It relieves students of having to give forced answers. Students are not punished for avoiding an item: choosing the ?-option renders a score of 0 points, while choosing the good answer scores 1 point, and choosing the false answer scores -1. Thus, the introduction of this new option is an attractive one for students: in the long run, such a system gives the same score as pure guessing, but in addition to that, it gives better feedback. Further, in the short run, the system is much saver than guessing, especially when the test constructors play their favourite game of formulating the items in such a way, that guessing gives a more than 50% chance on the wrong answer.

The block test being solely of formative nature (in later years, this situation was changed to some extent: in principle, the test remained purely formative, but for a small group of students who found themselves in a grey area between the white of success and the black of failure, it could become summative), without doubt the progress test is to be considered the utmost important instrument in the assessment system.

An outsider of our university would have a difficult job distinguishing a block test from a progress test. Both consist of items in a true/false format, the number of items being somewhat higher in the progress test: at least 250. The great difference is not a matter of format, but a matter of content: the items of the progress test are sampled from the whole cognitive domain, and represent the end objectives of the curriculum. Since the progress test is administered four times a year to all students of the faculty, irrespective of the year they are in, the test can best be conceived of as a kind of ‘repeated final examination’ (van der Vleuten, 1989, p. 14). Each student will participate in at least 24 progress tests, these tests
being repeated every 3 months and every time being made up of new items, parallel in content to the previous ones. This last aspect is achieved by sampling the items with fixed weights for the several disciplines, classified by a fixed blueprint.

Confronted with a final examination, one would not expect much from a freshman. A second year student would probably do a better job and, as we hope, the student just before graduation is at her or his best. Clearly, this is a vintage model, in which the expected productivity grows with the age of the vintage. The model also makes it evident why there is a need for a question mark option in the item format: if you expect the youngest vintage to reach a very low score, especially at their first test (ideally a zero score; when the test is constructed in such a way that it truly measures the contents of the blocks), it would be beyond all reason to expect them to go all the way, patiently guessing every item. Therefore, in the progress test even more than in the block test, forced guessing should be circumvented.

Since we have a vintage model with growing expected productivity, the obvious norm would be one that grows with the vintage. This is indeed the case, however in an indirect way. Grading is based on a so-called norm-referenced perspective of the test scores. This means that the relative score counts: the score of any student is compared with scores obtained by other students from the same vintage. (This contrasts with the so-called domain-referenced perspective, in which only the absolute score matters, and in which the performance of the fellow students is not relevant in the grading decision.) A good student scores above the average of her or his vintage; a bad student scores below that mark. Yet this is only one part of the story. The other part is, not surprisingly, the growth in the average score of students over time. Having to do with a relative norm, combined with a growth in average knowledge in her or his vintage, confronts the individual student with the need to perform better all the time (from an absolute perspective), in order to pass the progress tests.

The Paradigm Shift And Before

Already from the first year of its existence, there were differences between our assessment system and the ‘classical one’, especially with regard to the progress test. The main difference was related to the level of the test. Whereas the classical progress test aims at the final level, we started to test at an intermediate level. At that moment we did not have a real choice: only the first year curriculum had been developed, all the other parts being in a drafting stage. In other words, choosing the complete first-year program as the level of testing was more a necessity than the outcome of an unrestricted choice.

From the medical school we learned that testing at the final level has the disadvantage of a very low discriminative power in the first years of the study. This is not surprising: the freshmen are expected to master only a very small portion of the items, and, as a consequence of different individual learning paths, these portion differ from student to student. Yet in such a situation the sample error caused by sampling only a limited number of items from a very large cognitive domain, cannot be anything but large. This in turn implies a low discriminative power of the test within the group of these students. In the medical school, this situation is not very problematic. All medical students are bound to our university, due to legal restrictions to switch from one university to another (there is a so-called ‘numerus fixus’, combined with a selection committee). For our studies however, these restrictions do not exist, which means that any student who passes her or his ‘propedeutic’ (first-year) examination in Maastricht, is allowed to
continue this study at another Dutch economic faculty, or vice versa. This circumstance
necessitates a highly reliable examination at the end of the first year; passing the 'propedetic
examination' enables the student to continue her or his study elsewhere, and must, for that
reason, be a guarantee that the student is capable of finishing this study. Such a guarantee can
only be provided by an examination that is highly discriminative, which implies an examination
at a lower level than the final one. Being confronted with such a different legal context, one of
the first deviations with regard to the classical progress test was the inclusion of a large portion
of items at the first-year level.

A further step to reform the progress test into a solid guarantee for being capable of finishing
the study of economics, either at our own university or one of the other Dutch faculties,
concerned the grading system using the scores of the progress test. The medical school chose
for a group-referenced interpretation of the test score: the student's score is compared with the
score of other students of the same year group. The relevant score in this grading process is not
the score on one progress test, but to restrict the impact of outliers, a moving average of the
latest three scores on the progress test. We took over that grading system, but supplemented it
with an additional requirement: a criterion referenced score interpretation. In any of the four
progress tests taking place in the first year, students had to achieve an absolute minimum level.
In practice, this had to be the last progress test of the year, the test itself being composed of
items sampled from all first year blocks. Rather soon it appeared that for a large majority of our
students, this new requirement was the binding one. Which implies that we in fact changed the
assessment system having a progress test with a relative norm, formulated in terms of a moving
average of test scores, into a system with an absolute norm for one specific test: the last
progress test of the year.

Even then, the reform wasn't finished yet. In the medical school, the domain of 'common
knowledge' is quite impressive when compared to the knowledge that is rather specific for one
specialisation within the medical school. When we consult our doctor, and especially in the
situation we find ourselves in sickbed, we expect her or him to be an all-round professional,
being able to discriminate between the symptoms of pneumonia and bronchitis, but also
competent to recognise a concussion. However, we don't expect a political economist to be a
good advisor in labour conflicts, nor do we expect an econometrician to be a good negotiator on
export contracts. Implicit in this difference of expectations is the view, that the cognitive
domain of an economic faculty is much more diverse than that of a medical school. Looking at a
general economists, econometrician or business student, the common knowledge they share
refers to a much smaller domain than the specific knowledge we expect them to master. To
account for so many diverging specialisations as one typically finds in the study of economics,
we had to introduce many different progress tests, one for each specialisation.

The several reforms necessary to adapt the progress test to the context relevant for the
economic faculty, meant a drastic change of the assessment system. Even then, we were not
satisfied with the new system, and especially with the scoring model in relation with the item
format. To correct for guessing, in the true/false format choosing the false option is punished
with a score of -1, whilst the student who avoids answering the item, earns a neutral score of 0.
This would be a fair scoring model if the students who choose the false answer have less
knowledge than the students who choose for the question mark option. But this assumption
does not hold. Repeatedly we investigated the mean knowledge level of students choosing the
false option and students choosing the ?-option, and nearly always the first one exceeds the last
one. A result that does not surprise: a student who avoids to answer an item, probably has no
knowledge at all on that subject. A student who chooses the false option, probably has some knowledge, at least enough to think that she or he masters the subject. But this knowledge appears to be just partial knowledge, and the option chosen to be the false one. (Remark: this situation is not unique for the true/false format. Also the more well known multiple choice format finds itself in this same position. Students with no knowledge on the subject will guess, and earn the guess score. Students with partial knowledge are attracted to one of the false alternatives, and earn a zero score.)

The rather complex assessment system, the technical scoring problems mentioned above, and the feelings the format of fixed-response items were not suited for measuring problem-solving skills, all together led to the decision to abolish the progress test, and choose as a substitute the OverAll Test.

**Overall Test**

In the assessment system of Maastricht’s economic faculty, two separate tests exists to measure different cognitive behaviours within the same cognitive domain. Using Bloom’s taxonomy, the Knowledge Test has as its primary aim to test at the level of knowledge and comprehension, whereas the objectives of the OverAll Test refer to the levels of application, analysis, synthesis, and evaluation. Other objectives of the OverAll Test are to measure student’s problem solving skills, and her or his competencies in scientific reasoning. Shortly, the OverAll Test measures the creative competencies of the students, whereas other tests, such as the Knowledge Test, mainly tests the reproductive competencies of the student.

It is difficult to enforce creativity, especially under the conditions that are typical for an examination: many students gathered in a large sports hall, not only competing for high scores, but also competing for nervousness, making together so much rumour that even the execution of a routine action becomes rather awkward. To avoid the negative impact of these unattractive conditions, and to allow for an extension in time of the creative moment, the items of the OverAll Test are all based on scientific articles that are distributed among the students two weeks before the examination takes place. Together with study guidelines, that form an important clue for the items the students can expect at the exam. The combination of articles, study guidelines, a two-weeks preparation period without any other duty, and the open-book character of the test, create strong resemblance to a take-home exam. The important difference being the better guarantee we get on the intellectual ownership of the answers the students give. Although the study guidelines are an important clue for the items, they aren’t identical to the items, a fact that strongly reduces the risk of unwanted forms of co-operation.

The OverAll Test consists of both fixed-response format items (true/false items) and items with constructed-response or essay format. The latter format is however dominant, both with regard to the contribution to the total score the students can achieve, as with regard the time the student use to answer the items of both formats. Since the discussion on the pro’s and con’s of using constructed-response or essay format items does not seem to be conclusive, we performed a factor analysis on all the item scores. The outcome was surprising, in the context of the surveys reported in several journals (see e.g. Walstad & Becker, 1994). We found two different latent factor, in to one latent factors in most other tests, such as the Knowledge Test. The true/false items were loading at the factor, almost without exception. At the other side, the essay items were loading on both latent factors. What did this fact prove? Although we did not
know anything for sure, it was tempting to formulate the following conjecture: the first factor represents the basic cognitive behaviours of knowledge and comprehension, whereas the second factor represents the hierarchically higher levels of application, synthesis, and evaluation. Without being able to formally test this conjecture, we were able to do some additional research. We added to the scores on the different items of the OverAll Tests the scores on the different items of the Knowledge Tests. And again applied factor analysis to this enlarged data set. The outcome of the factor analysis did confirm our conjecture or, to formulate the conclusion more prudently, did not falsify it: the items of the Knowledge Tests, all of them being of true/false format, did indeed load on the same factor as the true/false items in the OverAll Tests. Which seems to indicate, choosing again a carefully-worded formulation, that in the context of a open-book test such as the OverAll Test consisting of items based on articles that the students can review in advance, constructed-response items and essay format items do measure different aspects of cognitive behaviour.

After this intermezzo on item formats and the testing of different categories of cognitive behaviour, I won’t spend more space on a further description of the OverAll Test, but refer instead to the next chapter by Mien Segers. And conclude this contribution by giving one typical illustrations of the OverAll Test, taken from.

Usually the level of articles in scientific journals as the American Economic Review is far beyond that our first year students can handle. Sometimes there are exceptions to this rule, and Battatjo ea. (1991) is one of the exceptions of such an exception. It is an article within the relative young branch of experimental economics, on rats and the axioms of the theory of consumer choice. It contains a description of an experiment designed to find conditions in which the Giffen phenomenon is present. These conditions aren’t so attractive; in fact they are so unattractive that we do not dare to ask human beings to participate in such an experiment. But rats cannot choose, and so the article contains many tables demonstrating the Giffen phenomenon present with several different rats. Tables that allow us to ask nice interpretative questions on these tables, on income and substitution effects that can be calculated from these tables, and so on.

References