Selected-Response Examinations
In A Student-Centred Curriculum

Dirk T. Tempelaar
University of Maastricht, the Netherlands

Key words: Assessment, Student-centred learning, Problem-based learning, Progress Test, True/false items, Scoring

1. INTRODUCTION

Multiple-choice and true/false items are popular formats in our examinations. In both traditional and more innovative educational systems, examinations using these two item formats form the backbone of most assessment systems. In fact: these formats seem completely self-evident, since many educational reforms appear to imply a change in assessment objectives, but not in its format.

In this paper, the self-evidence of multiple-choice, true/false and other forms of selected-response items is questioned for its use in innovative educational programs. Especially when the innovation is directed at achieving a student-centred system rather than a teacher-centred one, or stressing the importance of permanent knowledge rather than short-term knowledge, or implying assessment at end-level rather than intermediate level, the use of constructed response examinations and standard scoring rules is suspect.

We will give examples of the shortcomings of multiple-choice and true/false items in such innovative programs. The role of partial knowledge, and the possibility to discriminate correctly between students with partial knowledge and students without any knowledge, play key elements in this analysis. It will be argued that traditional scoring rules can imply an intransitive ordering, and thus an invalid discrimination, for broad categories
of items (the trick item being the most famous example of them). The paper gives a classification of different types of items that cause such an invalid discrimination. In addition to that, instruments are sketched to detect invalid items, and recommendations are given to avoid the construction of this type of items.

2. FROM NUMBER RIGHT TO FORMULA SCORING

The shift from 'teaching by teachers' to 'learning by students' is a common element in many educational reforms. Problem-based learning is without doubt a fine example of an educational system in which the learning activities of the students constitute the core of the educational program, but at the same time it is not the only representative of the shift of control in the educational process from teacher to student. For that reason, I will try to generalise our faculty's experiences with assessment in a problem-based curriculum to the more broad issue of assessment in a student-centred curriculum. In fact, most of what is discussed in this contribution also relates to assessment in a traditional curriculum. However, some symptoms manifest themselves much more pronounced in a student-centred curriculum than in a traditional one. To get a more precise impression of these unattractive symptoms of selected-response examination, we need a short introduction into the dominant habits of norming such examinations: the correction for guessing.

Which students take the correct option in a true/false or multiple-choice item? In the first place: those students who possess sufficient knowledge to give the correct answer. For these students, we can view the process of making this one item as a deterministic process: they know the answer for sure, and will pick the correct option with complete certainty. But the rational student will not skip any item if he has less than complete certainty, but at least some positive chance on earning credit for that item. Just imagine the student who is completely ignorant of the knowledge tested in the item. The best he can do is guess, and have a 25% chance to earn the credits of a correctly answered four-option multiple-choice item, or even a 50% chance of guessing the correct answer on a true/false item. Examiners faced with this problem of students, who merely guess the right answers to some of the test items, will be tempted to correct for this effect. Test theory tells us that this correction for guessing isn't too big a problem. Just leave the 'number right scoring norm', where we only count the number of correct answers, and give no credit for omitted items, to replace it with 'formula scoring'. This formula scoring can take on two different, but mathematically equivalent, forms. The first form is to reduce the individual score by the
amount that one is expected to guess correctly. When guessing on a k-option multiple-choice item, on average we will guess correctly in 1 out of k cases, and guess incorrectly in (k-1) out of k cases (since there is 1 correct option against (k-1) incorrect options per item). So we can find the expected guess score by dividing the number of wrongs, W, by (k-1). Subtracting that number from the number of rights, R, gives the so-called 'rights minus wrongs correction':

\[ X_c = R - \frac{W}{k-1} \]  

(1)

The other way to calculate corrected scores \( X_c \) is to allow for omitting items, and to credit omitted items with a partial score, that equals the expected score of guessing: \( \frac{1}{k} \). We then arrive at the following formula, with O being the number of omitted items:

\[ X_c = R + \frac{O}{k} \]  

(2)

Correction for guessing, in one of the two forms mentioned above, is a broadly applied technique. And in a student-centred curriculum using selected-response examinations, it really is a must, as I shall demonstrate in the next chapter.

3. ASSESSMENT IN A STUDENT-CENTRED CURRICULUM

A major problem in the design of student-centred curricula in general, and problem-based curricula in specific, is to circumvent the creation of a 'hidden curriculum'. If assessment of student achievement is not congruent with the educational principles, tests and examinations will become the dominant factor that drives the student learning, and not the educational objectives. An important, probably the most important, educational objective of student-centred learning is the principle of self-directed learning: students defining their own learning objectives during the tutorial meetings. Although the final objectives of the program are fixed, the students have the freedom to determine along which path they are going to achieve these goals. They use different learning resources, study subjects in a different order, and so on. As a consequence, learning activities and thus the knowledge acquired during a course are diverse, and assessment of student achievement cannot be directed at a uniquely defined, specific course
content. Instead, achievement tests in a student-centred curriculum are based on a much broader content, i.e. are related to the envisaged objectives of the program as a whole. By breaking the direct coupling between the content of the course and the content of the test, one can prevent a steering-effect from the test on the learning activities of the students. The Progress Test developed at the Maastricht University is an example of such a broad test independent of any specific course (see Van der Vleuten et al., 1996, and also Tempelaar, 1997). The Progress Test can be conceived of as a kind of repeated final examination. With final in the sense of: each Progress Test contains several hundreds of items, sampled from the entire cognitive domain of the curriculum. In addition to that, the Progress Test is a repeated examination, since it will be given at regular time intervals (e.g. each trimester) to all students of the faculty, irrespective of the cohort they are in. These repeated tests are made up of new items, parallel in content to the previous ones.

In such a Progress Test, students are confronted with many items of which they don’t know (and even cannot know) the answer. Those items can only be omitted. If the item-format does allow for a clear omit, e.g. by including a ‘don’t know’ or ‘question mark’ option, this is the obvious choice. If such an explicit option is absent, students are forced to guess. It will be clear that the inclusion of this omit option is much more elegant than forcing students to guess items that they, quite legitimately, do not master. However, whether an explicit omit option is available or not, this circumstance does not change much the problem discussed in the remainder of this contribution.

Because the cognitive domain to be tested in a Progress Test is so broad, there is a need to include many different items, which is only possible in case of a rather short item-format. This rather practical argument brought the architects of the Progress Test to opt for the true/false item-format that, supplemented with the option to omit, became the true/false item-format. In effect: the true/false format is identical (or in fact, encompasses) the two-option multiple-choice format.

A freshman is not expected to master many items, but a second year student already some more, whilst students in the final year of their study are expected to approach full score (or alternatively, to reach the same score as a reference group of graduated students). Bringing all test results of an individual student on successive Progress Tests together in one graph, one expects a gradually growing score over time: the student is expected to demonstrate some progress. This individual progress is compared with the average progress of all students in the same cohort, and the difference is used as input for examination decisions. However, these decisions are not
based on the result of one single test, but on a moving average of the results of several tests.

The essential feature of the Progress Test, to turn to the topic of this contribution, is the fact that students are confronted with many items they are not assumed to master. This characteristic is not unique for the Progress Test, but is also relevant for other forms of tests, as developed within student-centred curricula. Take e.g. a variant of the Progress Test, discarding the characteristic of repeated examination: a course test that includes items on optional subjects, to allow for individual learning paths. Students are expected to master e.g. one third of the subjects brought together in the test; however, which items belong to this 33% of required knowledge differs from student to student. One expects a student to answer those items that correspond to the subjects studied during the course, and to omit the items belonging to subjects that were not selected in the preferred learning path. Another way to phrase this conclusion is: the only way to reward different learning paths, and so to stimulate students to take their own responsibility in shaping their learning process, is to allow for different tests. An obvious solution to the problem of having different tests without abandoning the system of computerised, selected-response items, is to allow students to compose their own test, by answering a prescribed portion of the test items, and omitting the others. Having true/false items with an additional ?-option is one of the possible formats to achieve that.

4. TEST THEORY AND CONSISTENT SCORING MODELS

A test serves many purposes; the pass or fail decision is only one of them. Supplying the student with feedback on the learning process during recent courses is a good second, immediately followed by supplying lecturers or course organisers with feedback on the effectiveness of their efforts. To use test results to generate this type of information, the tests themselves have to meet several criteria. Test theory, or more specific: Item Response Theory (IRT), tells us these. In the most commonly used IRT models the following, related, assumptions are made (see e.g. Hambleton & Swaminathan, 1985, or Weiss & Yoes, 1990):

- unidimensionality: there is only one ability (latent trait) that accounts for the examinee’s test performance;
- local independence: examinee’s responses to different items in a test are statistically independent;
- speededness: examinees who fail to answer test items do so because of limited ability and not because they fail to reach test items;
• guessing: guessing does not occur.

It is not difficult to see that all assumptions are related to the assumption of unidimensionality. E.g., if the test is administered under speeded conditions, then at least two traits are impacting on test performance: speed of performance, and the ability measured by the test content. The same is valid for the last assumption: if guessing plays a role, the strict relationship between latent ability and test score is mixed up by a second one: chance, or testwiseness (or any other ability independent of the mastery of the test content) also determines test score. In the presence of guessing, or omitting items by choosing the omit option, models based on the assumption of unidimensionality collapse, and conclusions based on such models are not valid.

The IRT criteria for good items enable us to deduce a consistent model for test items. A convenient way to do that is by using a graphical representation of the requirement of consistency. This representation is called the ‘item characteristic curve’ or ‘item response function’ (IRF) and it plots the relationship between examinee's item performance and the ability that explains performance. Item Response Theory (IRT) requires this relationship to be a monotonically increasing function: examinees with higher ability are assumed to have a higher probability of a correct response. So IRT does constrain the IRF to be an increasing function of ability, but it does not determine the mathematical form of the IRF. In fact, alternative forms are possible, all consistent with the general framework of IRT, ranging from a simple linear function to non-linear functions as the logistic one. In Figure 1, several hypothetical examples are given of IRF's following logistic models with different parameters.

![Figure 1: Item response functions for five items](image-url)
When applying IRT, one has to choose one of these models, and subsequently has to estimate the unknown parameter(s) on the basis of the test results.

In this contribution, we will follow related, but somewhat different approaches that circumvent the tedious estimation of the parameters of the IRT model for each item. Thanks to the large number of observations on each item, we can calculate the 'empirical IRF' by averaging scores of students within different categories of ability, without the need to estimate any parameter. As an even simpler second approach, we distinguish the several answer categories that correspond to the item format used (two categories, correct and incorrect, in case of a true/false item, and three categories, correct, omit and incorrect, in case of a true/false item) and subsequently calculate the average ability level of all examinees opting for that answer category. We can substitute the examinee's total test score for the unknown ability level, and find an 'inverse IRF' without the necessity to explicitly estimate some model parameters. In this alternative set-up, the assumption of a monotonically increasing relation between ability level and the probability to give a correct answer, translates into a 'transitivity assumption':

- a better answer (given the ordering: correct is better than incorrect, and omit, if present, is in between) on a specific item corresponds to (i.e. can only be explained by) a higher ability level, and
- examinees with the same answer possess the same level of ability.

We can use this 'inverse IRF' plot and the presence (or not) of the transitivity property as instruments to judge the quality of individual items. But before doing that, we will investigate several scoring rules on their consistency.

5. **TRANSITIVE ITEMS**

In order to study scoring rules, we substitute the relevant score for each of the answer categories. The following figures plot the relationship between score (horizontal axis) and ability (vertical axis) for two out of three different cases:

- number correct scoring without guessing,
- number correct scoring with guessing, and
- formula scoring (number correct minus number incorrect) with omitting.
The most simple (and less realistic) case is the true/false format in combination with number correct scoring. This case refers to a complete dichotomous world: with respect to any item and any student, the state of knowledge is at one of the poles perfect, or completely incorrect. As a result, there are only two answer categories, all students possess knowledge about all items, and pick an answer on the basis of the state of their knowledge. Their knowledge is either correct, in which case they pick the correct answer, or incorrect, resulting in the incorrect answer. Having no knowledge about a specific item doesn't come into this picture: it is excluded by assumption. As Figure 2 demonstrates, the model is consistent: it obeys the transitivity constraint.

![Figure 2: Inverse IRF for number correct scoring without guessing](image)

The situation gets more realistic when we take guessing into account. There are now three ability levels: correct knowledge, incorrect knowledge, and no knowledge. However, there are still two answer categories: the correct answer and the incorrect one. Students without any knowledge are forced to guess, and are evenly spread over the categories correct and incorrect answer. We immediately recognise that in this case the transitivity constraint does not apply. In both answer categories, we find students with different ability level. And students with an identical ability level (no knowledge) are spread over different answer categories. Some are lucky in guessing, and get a positive score; others have less luck, and get no credit at all.

However, the problems caused by the phenomenon of guessing go even deeper than making the model inconsistent from an IRT point of view. In order to apply formula scoring we have to make an explicit assumption about the knowledge levels of students taking the different options. In the mechanism of formula scoring the "random-guessing model" is assumed: students whose knowledge tested in the item is not complete, chose one of the options mentioned in the item at random. But if this assumed model
does not apply, then formula scoring is not appropriate. In the discussion of
the third case, we will give an example of that.

Switching to this case, we again distinguish the three ability levels
correct knowledge, incorrect knowledge, and no knowledge, but apply
formula scoring, and include an explicit omit option. Therefore, the model
counts three different ability levels and also three answer categories, which
opens the prospect, by making appropriate assumptions, to recover the
transitivity property. If we indeed can make the appropriate assumptions, the
'Inverse IRF' would have the form graphed in Figure 3.

![Figure 3: Inverse IRF for formula scoring with omit option](image)

Since formula scoring credits the correct answer, and is neutral towards
the omit option but penalises the incorrect answer, it is important to realise
what, rather strong, assumptions are needed to turn this scoring rule into a
consistent model: the ability level of the student omitting an item is in
between the ability level of the students giving the incorrect, and the correct
answer. So, consistency requires that the state of no knowledge is preferred
above the state of incorrect knowledge (and both of these are dominated by
the state of correct knowledge). In the next section we will see that this
assumption is not met for the majority of items in our Progress Tests.

But before turning to the empirics, we will try to develop an intuitive
explanation for the discrepancy between actual item characteristics and the
assumptions behind the item model. In the item model three different,
hierarchically ordered ability levels are distinguished: incorrect knowledge
(say ability level: -1), no knowledge (ability level: 0) and correct knowledge
(ability level: 1). The crucial assumption of the item model is that the state
of 'no knowledge' is not the bottom of the knowledge scale, the pole
characterising the lowest possible ability level, but is the centre of this scale.
So according to the item model, there are other, really detrimental states of
knowledge, much worse than having no knowledge at all, that ought to get
severely punished. In other words, there has to exist something like 'anti-
knowledge. We found it nearly impossible to find convincing examples of such knowledge, since in most situations the absence of any knowledge is really the bottom (but remark: the item model forces the item writer to specify the body of such knowledge for every item designed, to let the scoring system be consistent and fair).

However, a much more intuitive and common interpretation of incorrect knowledge is that of 'partial knowledge'. In the context of partial knowledge, a student chooses the wrong alternative as a result of making a calculation error, taking no notice of a relevant piece of information, making an essential error in reasoning, and so on. As will be clear from these examples, partial knowledge is in no way the lower extreme point of the ability scale. In contrast, partial knowledge is somewhat of the centre of this scale (say, ability level: 0), with correct, complete knowledge as the upper pole (ability level: 1) and no knowledge as the lower pole (ability level: -1). But if this partial knowledge model applies, then the use of formula scoring based upon the random guessing model is not consistent and thus unfair. In fact, a consistent and fair scoring rule in the partial knowledge model is to treat the incorrect answers neutral, and punish the omits. Which clearly is not a stable scoring rule, since no student would use the omit option under that regime.

6. INTRANSITIVE ITEMS

The inconsistency that arises in the 'partial knowledge model' is sketched in Figure 4. At the same time, Figure 4 gives the empirical inverse IRF of the average item in our Progress Test. Having calculated this inverse IRF for each item in many Progress Tests, we find that a large majority of them does not fulfil the transitivity property. The deviations can be classified in two categories, one being quite more severe than the other.

![Figure 4: Inverse IRF for weakly intransitive item](image-url)
The first category is called the 'weakly intransitive item'. Compared with the transitive item, the ability levels of the students giving the incorrect response, and those omitting the item, are reversed: see Figure 4. Students choosing the omit option are the low-ability students, whereas the incorrect answer is given by the average-ability students.

The 'strong form' of intransitivity refers to the cyclical reversion of the ability levels of the students giving the correct answer, those choosing omit and those giving the incorrect answer. Its empirical inverse IRF is graphed in Figure 5.

![Figure 5: Inverse IRF for strongly intransitive item](image)

The attractiveness of transitivity as an indicator of the quality of test items is its conceptual and computational simplicity: it is not much effort to determine if an item is transitive or not, and in the latter case, if it is weakly or strongly intransitive. In order to find these qualifications, we calculate the average ability level, approximated by average test score, of students in each of the categories correct answer, incorrect answer and question mark. This exercise is repeated for all items. If the average ability level of the students giving the incorrect answer exceeds the average ability level of the students giving the correct answer (ability(incorrect)>ability(correct)), we call the item strongly intransitive. If either the average ability of the students omitting the item exceeds the average ability level of the students giving the correct answer (ability(omit)>ability(correct)), or the average ability of the students giving the incorrect answer exceeds the average ability level of the students omitting the item (ability(incorrect)>ability(omit)), we call an item weakly intransitive. The remaining category contains the transitive items (ability(correct)>ability(omit)>ability(incorrect)).

Using data from several Progress Tests, we reached the conclusions:
- The number of intransitive items is rather high: a vast majority (65%) of the items are weakly intransitive, an important minority (20%) are
strongly intransitive (annotating that a strongly intransitive item is by
definition also weakly intransitive).

- An item can be intransitive in one year-group, and transitive in
  another. Yet, the variation over the several year-groups is modest,
  with the general pattern that intransitivity is at its maximum in the
  first year-group. Which is not surprising: since all items reflect the
  end-objectives, the number of students having complete knowledge
  will be rather small. Students without any knowledge will omit the
  item or guess, and have a good chance to do better than students with
  partial knowledge.

7. **FIRST PARADOX OF THE PROGRESS TEST**

Why bother about intransitivity? The answer is extremely simple:
because intransitive items diminish the statistical reliability of the test. If
there are too many intransitive items in a test, we cannot be certain that
students with a high score indeed deserve that high score, are the better
students. This statement can be illustrated with a simple case. Assume we
design a test that only contains weakly intransitive items. This test is used to
compare the achievements of two different students: student 'zero' who did
not even try to start learning (so a nice example of complete absence of any
kind of knowledge), and student 'median' who masters 50% perfectly, and
the other 50% partially. In case of an essay test or any other test with
constructed-response items, student 'zero' is not able to write down any
sensible answer and will earn a score of 0, whereas student 'median' scores
75 out of 100. However, if we measure achievement with a test containing
selected-response items (such as the true/false items in the Progress Test),
if all items are weakly intransitive (which does not seem to be so far from
reality) and if the students are not capable of discriminating between
complete and partial knowledge, then both student 'zero' as student 'median'
will earn the same score: 0!

As a second case, take a test containing transitive and strongly
intransitive items in the ratio one to one. We use this test to compare student
'zero' with student 'genius': that exceptional student who knows everything.
Again, under the same assumptions, both of them will achieve the same
zero-score.

Intransitivity as studied above is an immediate consequence of the
selected-response format. However, in Progress Tests we cannot do without
that format. Which leads to the following
first paradox of the Progress Test:
Any student-centred curriculum demands for tests that are final in nature and are not directed at specific course contents. Such tests contain many items sampled from a broad cognitive domain, to allow for legitimate omits. For such a broad test, the item format has to be as short as possible: the selected-response format. But simultaneously: selected-response format items bring with them the risk of low reliability, particularly in the circumstance of high omit or guess rates.

8. WRITING ITEMS WITH DISCRIMINATIVE POWER

The next step in our item analysis was to search for common characteristics in the items showing one of the two forms of intransitivity. We found many common patterns, but most of them can be lumped together: the temptation of item writers to construct discriminating items. One should not be surprised by the discovery of that temptation: take an arbitrary book on item construction, and many recommendations to be found point in just that direction. E.g., Ebel & Frisbie (1991), one of the best known sources for authors of true/false items, give the following guideline (pp. 149-150): "The job of a test item is to discriminate between those who have and those who lack command of some element of knowledge, regardless of the type of interpretation to be made of the scores. Those who have achieved command should be able to answer the question correctly without difficulty. Those who lack it should find a wrong answer attractive." Specific means that can be used by authors of test items to achieve this goal are: "word the item so that superficial logic suggest a wrong answer; make the wrong answer consistent with a popular misconception, use phrases in false statements that give them the 'ring of truth'." These discriminating items, or trick items as some would call them, attract the students with incomplete or partial knowledge to the wrong answer. However, the plausible but wrong options have no hold over students with a complete lack of knowledge: having too little knowledge to get possibly attracted by any alternative, those students simply omit the item. As a result, the item becomes weakly intransitive: although partial knowledge surely goes beyond complete lack of knowledge, the score is reverse: whereas partial knowledge gets punished, lack of knowledge is credited by a neutral score. Remark that this also applies when there is no explicit omit option, and students with a complete lack of knowledge are forced to guess. Being unable to appreciate the attractiveness of the distractors, students without any knowledge will make a pure guess,
and have a 50% chance to find the correct answer, whereas students with partial knowledge are attracted in the wrong direction, and will end up with far smaller chances to find the correct option. And thus we have derived the:

**second paradox of the Progress Test:**
the better the authors of test items comply with the recommendations of test theory, the more intransitive items they design and the less the statistical reliability of the test.

9. **ITEMS AT AN INTERMEDIATE LEVEL**

Weakly intransitive items can be explained by authors in search for discriminative power. To explain strongly intransitive items, we need another phenomenon: items at an intermediate level. The nicest definition of this case probably can be found in the following example: see Figure 6.

<table>
<thead>
<tr>
<th><strong>statement:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Under the gold standard, the price of gold is constant.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>clue:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
</tr>
</tbody>
</table>

*Figure 6: Item at intermediate level*

The item tests the essence of the gold standard (or any other metallic standard): gold is defined to be the denomination, which implies that its price is constant, by definition. However, the clever students are aware of the technicalities of a metallic system: the price of gold does fluctuate around the central gold parity between narrow margins, whereby the lower gold point is determined by the costs of minting gold, and the upper gold point by the costs of melting coins to bullion. So these more or less ‘gifted’ students (as compared with their companion students, but also with the end objectives of the curriculum) choose the false-option arguing: an essential feature of a metallic standard is that minting and melting take place; this will only happen as long as the price of the metal is allowed to fluctuate around parity.

Items of this intermediate type inevitably result in ‘strong intransitivity’. The (very) high-ability students choose the wrong answer, the average-ability students pick the correct answer, since the item is exactly at their
level, whereas the low-ability students omit the item. As can be seen from Figure 5, a cyclical interchange of the credits achieved by the students as a function of their ability level takes place.

Although the item is called intermediate, it is important to realise that this phenomenon will not only occur with items that literally are at an intermediate level, but also with items that indeed are final, but for which part of the students have a surplus of knowledge, e.g. achieved by reading higher level textbooks than their companion students. It is exactly this freedom of students to design their own learning path that makes writing test items such a difficult job.

10. ASSESSING PROBLEM-SOLVING CAPACITIES

Since problems and the students' ability to solve them play such an important role in our curriculum, it is an obvious challenge to incorporate this type of competencies in the Progress Test. So a majority of items go beyond the level of asking for some element of factual knowledge, but confront the student with a short but possibly complicated case. Solving this case will require both the application and the synthesis of skills and several pieces of knowledge. Such 'composite' items are very sensitive for intransitivity: a little error in one link of a long chain can lead to the incorrect answer, whereas lack of any knowledge will not. Which brings us to the:

third paradox of the Progress Test:
the better the authors of test items succeed in writing items that combine the assessment of different levels of knowledge, such as recognition, application and synthesis, the larger the risk for intransitivity and the less the statistical reliability of the test.

If such a composite item furthermore is worded in the now well known discriminatory way, we get a real trick item, characterised by strongly negative average correct-minus-incorrect scores and/or negative correlations.

11. CONCLUSIONS

Discriminating items, and items at an intermediate level, as examples of intransitive items of the weak and strong form, constitute the Scylla and Charybdis of student assessment in a student-centred curriculum. The freedom of students to follow diverse learning paths asks for broad tests,
covering many different subjects. This goal constrains in turn the format of
the test: it ought to contain selected-response items, and moreover, the item
format has to be short: true/false, or multiple-choice with a restricted number
of alternatives. The item format should allow for an omit: part of the items
will refer to optional knowledge.

Having fixed the test format, the item writer is really in the position of
carefully manoeuvring between Scylla and Charybdis. One can avoid the
risk of an item becoming strongly intransitive, by formulating discriminating
distractors. This will, however, increase the chance of weak intransitivity.
Reversibly, by choosing more neutrally formulated distractors, there is an
increased chance that the high-ability students with a surplus of knowledge
will be misled.

After several years of experiments with the Progress Test, this form of
student assessment was replaced by the OverAll Test in our problem-based
curriculum (see Tempelaar, 1997). Several considerations induced faculty to
this educational reform; the hard job to write adequate and reliable test items
being one of them.

REFERENCES

Ebel, Robert L. and Frisbie, David A. (1991): Essentials of educational measurement, 5th
Hambleton, Ronald K. and Swaminathan, Hariharan (1985): Item Response Theory:
Tempelaar, Dirk T. (1997), Congruence Of Assessment And Instructional System: The Case
Of Problem-Based Learning, This Volume, pp. 197-212.
experience with progress testing in a problem-based curriculum, Medical Teacher, vol. 18
(2), 103-109.
Weiss, David J. and Yoes, Michael E. (1991), Item Response Theory, in: Advances in
Educational and Psychological Testing, ed. Hambleton, Ronald K. and Zaal. Jac. N.,