AN ALTERNATIVE WAY OF USING MULTIPLE-CHOICE QUESTIONS

Frans Geurts

An Alternative Way of Using Multiple-Choice Questions

Frans Geurts
Wageningen University
Agrotechnology and Food Sciences
Laboratory of Physical Chemistry and Colloid Science
P.O. Box 8038
6700 EK Wageningen
Telephone+31 (317) 484297
Fax +31 (317) 483869
E-mail Frans.Geurts@fenk.wag-ur.nl
Internet www.fenk.wag-ur.nl

Abstract

At Wageningen University at the chemistry department we use an exam system with pre-printed answers (de Keizer, 1981). The exam is made up of a number of items. Each item is made up of a stem and a pre-printed amount of alternatives. There is no limit to the amount of alternatives per item. These vary between 2 and 40. We get these alternatives by following a number of solution routes from the stem. The chosen solution routes are taken from the different ways in which students answer such questions. Mostly all the answers (good, less good, wrong) we get in this way are presented to the student. Students must answer the questions/problems as if they were open questions. The exam consists of many problems designed in this way and alternatives calculated in this way. The student marks his answer on a pre-printed answer form. The correction of the exam takes place with the help of a key file in which every answer can be marked. The full amount of marks is awarded for a correct alternative. Wrong answers get zero marks. Some alternatives are also awarded (from 0.5 to nearly the maximum). The amount of marks received for an alternative is corrected with a guessing score that has been determined by the sum of marks to be awarded and the amount of alternatives. This is therefore a special form of weighed, forced-guessing, multiple-choice examination.

With the help of a computer program the answer form of each student (read with the help of an OCR apparatus) is compared with the key file. It takes a few minutes to get the results of up to 1400 students and the analysis of items, results of the students and the quality parameters are available from the computer program

The exam is composed by use of items already programmed in Excel. The solution routes of all items are calculated in these Excel documents. The items can be used as a source: changing a few parameters of the source item gives a new item with corresponding new answers. Quality parameters (difficulty, discrimination) of the source item are known and are a good estimate of the quality parameters for the “new” item.
Each Excel document is linked with a Word document. Pasting selected items in one Word document makes a new exam. Again a special computer program is used to finish the job.

Keywords

Question types
Use of IT to underpin the overall assessment process
Question storage in databases/question banks
Quality assurance issues

Introduction

At the Laboratory of Physical Chemistry and Colloid Science of the Wageningen University (WU) a specially developed test system is used among other things for the examination General and Physical Chemistry.

Designing an exam for General and Physical Chemistry

The exam is made up of a number of items. Each item is made up of a stem and a pre-printed amount of alternatives.

I (3.0)
Given the pKa of acetic acid is 4.74.
The calculated pH of a 0.1 M solution of acetic acid is:

1.1	2.12	1.5	4.74	1.9	11.13
1.2	2.37	1.6	7.00	2.0	11.63
1.3	2.87	1.7	9.26	2.1	11.88
1.4	3.12	1.8	10.88		

Figure 1, Example of an item

There is no limit to the amount of alternatives and the amount varies between approximately 2 to 40. The lecturer follows a number of solution routes from the stem. The chosen solution routes are taken from the different ways in which students answer such questions. Extensive experience of the different lecturers with this system makes sure that nearly all the solution routes, and therefore nearly all possible answers and alternatives, can be offered. Students must answer the questions as if they were open questions and their answer can be found in the pre-printed alternatives. The amount of

awarded marks is defined for every question. This varies between 1 and 8. For a correct alternative the full amount of marks is awarded. Wrong answers get zero marks. Some alternatives are also awarded (from 0.5 to nearly the maximum). The amount of marks received for an alternative is corrected with a guessing score that has been determined by the sum of the amount of marks to be awarded and the amount of alternatives. It is therefore a special form of a weighed, forced-guessing, multiple-choice examination.

The items used in previous exams are classed according to subject, including analysis and possible remarks and then stored, including their alternatives, stored in a computer in a special text-processing format, the @-format.

@ question @ points
Given the pKa of acetic acid is 4.74.
The calculated pH of a 0.1 M solution of acetic acid is:

@0100	2.12	@0500	4.74	@0900	11.13
@0220	2.37	@0600	7.00	@1000	11.63
@0330	2.87	@0700	9.26	@1100	11.88
@0400	3.12	@0800	10.88		

Figure 2, Example of an item, @ -format

Based on data in the specification table, the examination is composed of items (revised or not) from the available item bank and newly devised items.
A number of lecturers devise questions for a certain block (subject) and then, after consulting with the composer of the exam, these questions are combined to a balanced amount of questions distributed over the subject matter. Based on certain criteria the total examination is then assessed by other lecturers than the composers. The time span of the exam, the amount of marks per question, the time of announcing the results, i.e. all the data of interest for the student are then made known. It is attempted to group the items with a rising level of difficulty into 5 different blocks.
The final exam is put on paper and multiplied.

Correction of the exam

The answer form is read with the help of OCR apparatus. The data are then imported into the examination correction programme using the key file (*figure 3*)

Exam October 99

Line	Question	Number	0	1	2	3	4	5	6	7	8	9
1	IA	10	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
2	IA	3	3.0	0.0	0.0							
3	IB	10	0.0	0.0	0.0	1.0	0.0	0.0	3.0	0.0	0.0	0.0
4	IB	2	0.0	0.0								
5	IIA	6	0.0	0.0	3.0	0.0	0.0	0.0				
6	IIB	10	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	IIB	2	0.0	0.0								
8	IIIA	10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.5	0.0
9	IIIB	10	0.0	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0	0.0

10	IIIC	10	0.0	3.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	IIID	10	0.0	0.0	0.0	0.0	4.5	0.0	0.0	0.0	0.0	0.0
12	IIID	10	0.0	0.0	0.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0
13	IV	10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0
14	VA	10	0.0	0.0	2.5	0.0	3.5	0.0	0.0	0.0	0.0	0.0
15	VA	5	0.0	0.0	0.0	0.0	0.0					
16	VB	4	1.5	0.0	0.0	0.0						
17	VC	10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	VC	5	5.0	1.5	1.5	0.0	0.0					
19	VD	3	0.0	2.0	0.0							

Figure 3. Part of a key file

Psychometric analysis

*Item analysis

-Level of difficulty

In literature the level of difficulty p (Dousma, 1989) has been defined as the amount of students who have chosen the correct alternative, divided by the total amount of students. For the examination-processing programme it is the percentage of students who have chosen the correct alternative. A new level of difficulty has been imported because in the used examination system it is also possible to obtain marks for less correct alternatives. This is called the pm and is defined as the product of the amount of students who have scored a certain alternative of an item multiplied with the amount of marks awarded to that alternative, summed by all the alternatives of the item divided by the total amount of students multiplied with the amount of marks awarded to the correct alternative.

In equation:

$$p = \frac{\text{amount of students with a correct item answer}}{\text{total amount of students}} \left(\text{eventually multiplied with } 100 \quad (\%)\right)$$

$$pm = \frac{\sum_i (n_i \cdot point_i)}{max\_point \cdot \sum_i n_i (= n)}$$

These values can also be corrected for the guessing score of the item: p' and pm'.

The calculation of the p(m) and p(m)' values have been programmed in the original examination processing programme.

-Attraction and quality of the distracters/alternatives

In literature it has been described that the distracters that hardly score are not satisfactory and must therefore be eliminated. In my point of view this does not apply to the form of multiple choice as used for the examinations of General and Physical Chemistry. Several distracters are derived from thinking errors that do get marks awarded.

Besides this several distracters have been included which are absolutely wrong. These have been included to narrow down the estimating score. For a certain amount of questions there are so many distracters that it is perhaps useful to critically look at some of the distracters which no one scored, even if this alternative is awarded marks. Wrong distracters that have scored a lot should also be included in this critical analysis.

Many distracters are of good quality, that is they score reasonably well and are often awarded marks

-Discriminative ability and level of difficulty (corrected for guessing)

If we put pm' against r_{it} an idea of the quality of the examination and of the different amount of items is given. In the graph the marks 0.40;0.30 is given. Only a few items can be seen in the lower left corner. On the basis of this graph it can be easily seen which items should be selected for further analysis. Here it also holds that the experts should try to find out if and why a certain item obtains a certain score.

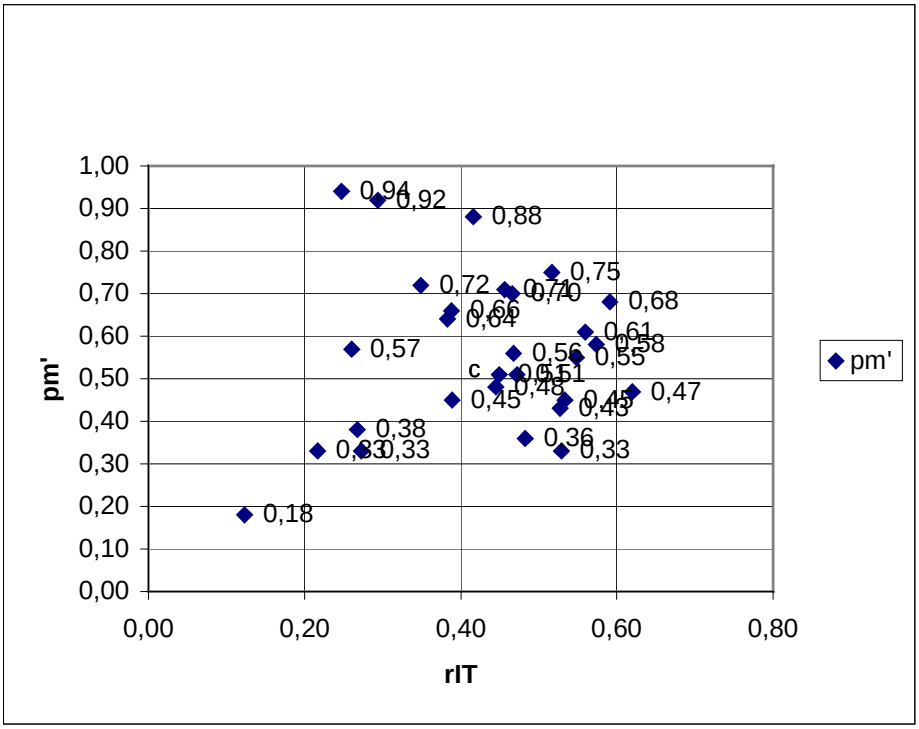

Figure 4.: pm' vs r_{it}, 1994, first year students

*Test analysis

-Validity

Using a specification table for a valid test is very important. The specification table used for the General and Physical Chemistry examination (divided into 5 blocks and per block a certain classification of awarded marks) could be extended by classifying the questions into levels. The simplest model that I could find in the literature was classification into 4 levels (Groot, de 1973a, 1973b, 1983). From investigations done with my colleagues it appeared that this classification was difficult in use.

-Reliability

For different examinations the reliability (r_{xx}) was measured; it is around 0.80 for all of the examinations.

-Objectivity of the score

The marking classification per item is determined beforehand and is given in the examination papers. The awarding of marks per item and giving marks to the several, partially correct, alternatives, is globally determined beforehand. Based on the first print out of the examination-processing programme the analysis of the examinations and the items can be looked at critically. There is of course some subjectivity in awarding points per item and per alternative. Once the marking classification has been determined the scoring is objective and definite.

Scoring and Marking

After analysis and necessary adjustments, the marks of the test are determined by a linear equation:

$$T = 1 + 9 * \frac{(s-g)}{(t-g)}$$

Subsequently, after fulfilling the practical training demands, the following equation is used

$$E = 0.88\ T + 1.2$$

The final marks* are calculated.
T = pre-examination marks;
s = score student;
g = guessing score;
t = maximum score;
E = examination marks.

The examination marks are rounded without decimals and are then electronically sent to the central marking processing system.

*Marking in the Dutch education system is from 0 –10.

Individually extended exam (IVE)

For students who only just failed (i.e. with 5 marks) there is, under certain conditions, the possibility of taking an individually extended examination (Van den Berg, 1992). The extended examination is mainly aimed at the individual student's weakness (via analysis of the blocks). It is possible for the student to do a re-examination in their two, worst blocks. The procedure goes as follows; for every student who is eligible for an IVE a packet is composed with an overview of the score of the examination, the correction-key, the block distribution with a description of the subject matter, the blocks which will be tested again and some general study advice.

The IVE is also analysed in the manner explained above.

To stop incorrect study behaviour and standardization in distinction, a bonus/malus system is applied when assessing an IVE. That is that the IVE is treated as a separate examination with separate marks. If these marks, after rounding off, are sufficient (=5.5) then the final marks for the total course are 6. If the IVE- marks are 4.0 or lower, then the final marks will be 4. If the rounded off IVE marks are 5, then a 5 will be maintained as the examination marks.
Table 1 gives an overview of the IVE results.

Table 1. Overview of IVE results

year	total amount of participants	% participants who passed	% participants getting final marks 4
1992	50	44	20
1993	56	59	13
1994	47	43	34
1995*	80	51	20

*too ridiculous" clause used.

Processing speed

If a three hours exam is taken in the morning of day X, then the whole correction procedure must be finished in the afternoon of day X + 3, i.e. the students (in the last few years approximately 500) know which final marks have been obtained and if they are allowed to take the IVE.
So if the exam is on Monday December 21, then the results must be known on December 24 and the student will know if he can take the IVE, possibly in the beginning of January, just before the start of the new trimester. The student who seriously wants to participate in this examination must offer up his Christmas holidays.

Results of december examinations

In the table below you will find the percentage of students who have taken the December exam of the last four consecutive years. As a starting point marks of 4.89 are taken (with a re-evaluation for practical training the final mark is a 6, IVE is not included).

Table 2. An overview of the results of the December examinations

year	total amount of participants	% of participants who passed	total amount of A-level participants	% of A-level participants who have passed
1992	500	65.8	407	70.5
1993	503	47.1	348	55.2
1994	508	63.0	363	66.9
1995	561	38.7	426	42.8
1995, after correction*	561	49.9	426	54.2

*”too ridiculous" clause used.

From the table above it is evident that:
– the percentage of students who have passed varies from year to year
– the percentage of A-level students who have passed is higher
– the percentage of A-level participants varies from year to year.

Exam processing

From *figure 5*, the percentage of students who have obtained a score on a certain alternative and how many marks are awarded to this alternative can be seen in the printout per question of the test-processing programme.

```
Question IV      : 30 '*' is 508 students (100%)

  Altern.   pnt.  numb.  (%)

      13.0          5 (  0.98) |

      13.1          2 (  0.39) |

      13.2          0 (  0.00) |

      13.3          3 (  0.59) |

      13.4   4.0  182 ( 35.83) |***********

      13.5          7 (  1.38) |

      13.6   2.5   93 ( 18.31) |*****
```

```
    13.7   3.0    1 (  0.20) |
    13.8          2 (  0.39) |
    13.9   1.5    2 (  0.39) |
    14.0          0 (  0.00) |
    14.1          0 (  0.00) |
    14.2        105 ( 20.67) |******
    14.3         51 ( 10.04) |***
    14.4         47 (  9.25) |***
    empty         8 (  1.57) |
p      pm     p'     pm'
0.36   0.48   0.31   0.44
```

Figure 5.: print out per question

In an overview it can then be seen how every item has scored in relation to the average total score. The total guessing score of the test is obtained by adding up the guessing scores per item. If the guessing score of an item lies above 1.00, it will be brought back to 1.00.

The fact that it is known how every alternative is derived plays an important role in analysing and assessing the items.

However, time and time again it turns out that a certain item does not give the score the lecturers expected it to give. Afterwards it can (often) be analysed why a certain item scored differently to what was expected (high or low) or to why the test was made so well/poorly. Predicting the level of difficulty of an item/exam is apparently very difficult.

Design and result of an investigation among colleagues

Per year three examinations of General and Physical Chemistry are given. The December examination is the examination with the highest amount of participants and is given following the teaching period. To be able to analyse several aspects I have chosen several December examinations. I mixed up all the items of the years 1992, 1993, 1994 and 1995 (to avoid recognition) and then again randomly distributed them over different years in the five blocks also used for the regular exams. It could not be seen from the numbering from which year an item came.

I asked my colleagues who give the lectures of General and Physical Chemistry, to estimate for every item which percentage of students would give a correct answer for that specific item; the p-value. By students I meant the regular first year students who

were doing the course for the first time. The regular first year students are, as regards pre-knowledge, the most stable population and the given lectures are based on this pre-knowledge.
The results found agree with the results noted in literature (Mellenberg, 1971, Frijns, 1993): it is very difficult, if not impossible, to be able to correctly estimate beforehand, if an item scores well or badly. Nor is there an agreement amongst the colleagues about the level of difficulty of an item. To estimate Wijnen (1993) advises to separately ask the experts' opinion about the difficulty, and then to take the average of their results as a directive. It is absolutely dissuaded to hold a meeting about the estimated values. It is advised is to average the sometimes extremely different values.

Reaction Of Students

The chosen examination form sometimes gives critical reactions. This criticism is directed at the fact that some students make a calculating mistake somewhere, or fail to re-calculate to the correct end answer, and that for a "partial answer" no marks are awarded. We can partially meet this argument for the first criticism by using letters instead of numbers. For the students the level of difficulty of the items seems to increase in such a way that this cannot be a real alternative. An alternative for the second argument could be cutting the different steps into parts. For example in part A the numerator of a certain result of this part could be asked for and in part B the denominator. This probably means that the examination, as regards level, will become easier. Using letters could then again raise the level of difficulty. Estimating the level of difficulty of the item is hardly possible.

Maintaining Standard

The aim is, especially for the December examinations, to use a large number of new items in the examination.
The above research shows that by inserting these new items with an "unknown" level of difficulty, the percentage of passes is less predictive. This means that maintaining the standard for the different years and making it equivalent with other years is not really possible.

Developing a Test Service System

Composing an examination is very time consuming and especially calculating the many alternatives takes a lot of time. Giving examinations with only one correct answer per alternative is not a good choice for a number of reasons.

There is a lot of literature about the "quality" of the tests. At the moment a lot of attention is given to the use of computers and networks for giving tests. One line of thought is the flexibility of time (every student determines the moment and the place of the examination themselves) and the other is adaptive testing, where, based on the

answers by the student, the examinations are held at a more difficult or easier level, and thus also flexibility in examination time. However, flexibility means an enormous investment in the material to be developed (software, but certainly also items) and therefore probably a bit more stability in the education offered.

In the case of General and Physical Chemistry we started automating the generating of the questions. By using a spreadsheet a large part of the existing questions can be developed as a sort of template for the question, with re-calculation of the alternatives. Adjusting some of the parameters can develop a new question for the students. Furthermore characteristic data of the questions can be stored (subject, pm', D, date, etc.). The concerned template question can be stored in a word-processor in the already developed @-format. Using a specification table, where the set demands are translated into a composition of the examination, can now develop a kind of test matrix.
The advantage of this system is that it can be done gradually and that the development of the template questions can give more stability in the scores obtained by the students, at least that the students themselves will mainly cause the instability.

Discussion

More time must be spent on reporting the results of the examination and processing the psychometric data. The demands set for the items to be admitted into the question bank must be high and the items must be re-assessed after every examination.
However, after only a short time span, when many items have been developed in this manner, a part of the item bank can be given to the students for practice. Especially the (automatic) feedback to the students by re-calculated alternatives can be done without too much time investment.
Most of these costs of giving the examinations are determined by developing the items with their matching alternatives, and if these costs are looked at, then this way of examining can be economically interesting.

The software to be developed for this application can be "plainly" written by using existing programmes such as Excel and Word. In this way a very flexible system has been developed.

References

Berg, A.B.A., van den et al. (1992) Verslag project Individueel Verlengd Examen: Wageningen University

Dousma, T. en Horsten, A. (1989) Tentamineren: Groningen, Wolters-Noordhof

Frijns, P.H.A.M. (1993) Over structurering van beoordelingmethoden voor open vragen: Maastricht, PhD thesis

Groot, A.D., de en Naerssen, R.F. van (1973a) Studietoetsen, deel 1: Den Haag, Mouton, 2e druk

Groot, A.D., de en Naerssen, R.F. van (1973b) Studietoetsen, deel 2: Den Haag, Mouton, 2e druk

Groot, A.D., de en Wijnen, W.H.F.W.(1983) Vijven en zessen: Groningen, Wolters-Noordhoff, 10e druk

Keizer, A. de (1981) Bevredigend alternatief voor dwangbuis klassiek meer-keuzesysteem?; Onderzoek van onderwijs, 10, 2, 7-8

Mellenbergh, G.J.(1971) Studies in studietoetsen: Amsterdam, Psychologisch Laboratorium, Proefschrift

Wijnen, W.H.F.W. (1993) Beoordelen in het onderwijs in Berkel, H.J.M., van, en Bax, A.E. Beoordelen in het onderwijs: Houten, Bohn Stafleu Van Loghum