A comparison of a classical calculus test with a similar multiple choice test

H. J. L. KAMPS AND J. H. VAN LINT

A COMPARISON OF A CLASSICAL CALCULUS TEST WITH A SIMILAR MULTIPLE CHOICE TEST

1. INTRODUCTION

A number of departments of mathematics at Dutch universities experimented on a small scale with multiple choice tests, mostly for calculus courses for freshmen. Although the results were not bad (with a few exceptions) none of these departments saw convincing reasons to continue with the method. In 1969, a period when Dutch students tended to protest against practically everything, there was a movement at the Technological University Eindhoven criticizing the essay-test method used by the mathematics department and demanding 'objective' tests. This was the motivation for a more serious examination of this method by the authors of this report, who at the same time constructed two tests which would make a comparison of the methods possible. The results were published in a report [7] which seems to have had one useful aspect, to wit the fact that since its appearance we have heard no more about objective tests from the students (or anybody else). We have been requested to translate part of the report to make the results accessible for a larger part of the mathematical community.

2. A STUDY OF OBJECTIVE METHODS

To start our research we made a study of the literature on mathematical education, especially the parts on different methods of testing (cf. references). In this section we give a short survey of observations which we encountered which contributed to our understanding of the problem.

2.1. *Objectives of Education*

Bloom [2] divides the cognitive domain into 6 classes: knowledge, understanding, application, analysis, synthesis and judgment. With this subdivision as a guide the teacher must decide (cf. de Groot [4]) what is essential to the subject he is teaching for the group of students under consideration, e.g. are errors in the calculation serious errors or not, is it sufficient that the student shows that he understands more or less what the solution of the problem is or does one require an exact presentation. All authors agree on two matters: (i) it is the teacher who decides what the objectives are and what is a sufficient result on a test; (ii) the students should

Educational Studies in Mathematics **6** (1975) 259–271.

know these opinions. For this reason de Groot [4] prefers the word 'score' for the result of a test, leaving the decision on what is passing resp. failing as a separate problem. If one decides that a fixed percentage of the class will pass the test, then there is no problem. Otherwise one must say beforehand (!) which score is the lowest passing score. This is the difficult problem in designing tests and it is not made easier by using different methods of testing.

When discussing objectives of education one should also ask the question what the objective of a test is. Tests influence future education, they can be used to see if a course had the desired effect but also to see if a group of students has reached a level enabling them to understand subsequent courses. One must keep these things in mind when designing a test.

2.2. *Definition of Objective Tests, Applicability in Mathematics*

As was to be expected, the definitions in the literature differ. The most extreme definition is given by de Groot [4]:

> a test designed in such a way that the score of a student can be determined without the aid of a scorer who is a professional in the field, but e.g. by a computer.

Other authors call a test objective if 2 independent scorers obtain the same score for the tests. Other authors simply state that objective and multiple choice are the same or they claim (without any argument) that objectivity implies the multiple choice method.

In essay test examinations it often occurs that the answer given by a student is practically completely nonsense. However, most teachers tend to read such answers trying to find something which is worth a few points. It is this method of scoring which the advocates of multiple-choice testing call an injustice because it is not objective.

After studying sufficient literature, we had an idea of the average definition of the word objective. We came to the conclusion that a classical essay test (in calculus e.g.) with several small problems, for which it has been decided beforehand which answers will contribute (and how much) to the score, satisfies this definition. However, the literature states as objective only (1) multiple-choice, (2) multiple true-false, (3) matching, (4) short answer tests (e.g. ${}_0\int^{\pi} \sin x\, dx = \ldots$). In fact some authors do not even recognize the 4th possibility as objective.

Some quotations from the literature follow:
Ahmann [1]:

> It has been found that multiple-choice test items can be used at all class levels with the possible exception of the primary class levels. Other than the area of mathematics, the multiple-choice test item can be used successfully in all subject-matter areas which have verbal and mathematical aspects. Even in the case of mathematics, this type of test item

serves a useful function, unless there happens to be a heavy emphasis upon computational aspects of mathematics.

Husén [6], comparing essay tests with multiple-choice tests:

It is not claimed that the two types measure exactly the same thing, since they obviously do not.

Nunnaly [12] in an attempt to show how wonderful objective tests are, points out one of our main objections to using the method:

The truly skillful item writer can test almost anything with objective items. The reason why so many teacher-made objective tests do not get at more important parts of the content is that the teacher does not have the skill and/or the time to compose an excellent test.

De Groot [5] is more reasonable:

It is possible to translate a lot more questions, e.g. those requiring some thought or insight, into a precoded form than most Dutch critics of achievement tests believe, however not everything can be tested in this way. Independently solving a math-problem, the writing of an essay on some topic, ... ; one cannot simply replace these things.

Lindquist [10] gives a number of examples showing flaws of essay tests but he also describes the unfavourable effect of objective tests on the student's studying habits. He makes it quite clear that a good idea of the objectives of the course determines which method of examination should be used, and not slogans like 'justice', etc.

2.3. *Advantages of the Multiple-Choice Method*

From the literature we quote the following alleged advantages:

(i) They are 'fairer'; it is easier to justify the score to the student.
(ii) They are indispensable for the research and analysis of education.
(iii) Scoring is quick and efficient.
(iv) Decisions concerning passing or failing are better founded.
(v) It is possible to analyse the test systematically.
(vi) It is easier to keep the level of examinations constant.
(vii) The results of the test are more reliable.
(viii) One can ask more questions and thus cover a larger part of the subject material.

Remark: We believe that (iv) is nonsense (cf. e.g. the observations by de Groot mentioned in section 2.1).

2.4. *Disadvantages of the Multiple-Choice Method Compared to the Essay Tests*

The following list contains properties which some authors consider to be disadvantages of the multiple-choice method.

(i) Independent formulation, creativity and inventivity are not tested.

(ii) Sometimes multiple-choice questions are ambiguous, especially for the better students.

(iii) These tests have a negative influence on the education program. Students tend to learn facts and isolated pieces of information instead of general patterns and understanding. (Some authors claim that this is only so for poorly designed m.c. tests.)

(iv) Guessing occurs more often.

(v) Often one wishes to test whether a student can produce the correct answer, not whether he can recognize it.

(vi) Substitution or approximation often show which answer is correct (in mathematics).

(vii) Students expect the correct answer to be uniformly distributed over the alternatives a, b, c,... etc. This influences the final part of the test.

(viii) It is difficult to decide which score should be considered passing.

(ix) It is very difficult to design a good m.c. test, i.e. to find reasonable (but false) alternative answers.

Remark: It is our opinion that a number of these are not important, e.g. (ii) is probably not true for mathematics. However many of these disadvantages appear very significant.

Sections 2.3 and 2.4 show that it is difficult to compare the two types of tests

	essay test	m.c. test
design of the test	simple	difficult and time consuming
material covered	small sample	large sample
measurement of knowledge or understanding	both, mostly understanding	both, mostly knowledge
preparation by the student	ideas and principles	details
type of answer	own words	choice
guessing	no problem	difficult problem
scoring	difficult and time consuming	simple

and that much depends on what one wishes to measure. In this respect the following remarks from [1] are revealing:

For many years the essay test item has been the mainstay of paper-and-pencil testing. Its role today is often misunderstood. Rather than being totally replaced by the objective test item, as some suppose, it has simply released part of its function to the objective test item and is performing the remainder of it functions as effectively as before. That part which has been released is the part devoted to the measurement of the student's ability to recall information. It is in this area that the objective test item is extremely efficient. The informed teacher today is using objective test items primarily for this purpose and is using essay test items primarily for the purpose of measuring higher-level intellectual skills of the student.

2.5. *Constant Standards*

A major difficulty in designing tests, even for experienced teachers, is judging a priori how difficult the test is. Both students and teachers complain about tests which turned out to be much harder or easier than last years' test, etc. Authors favoring the m.c. test sometimes claim that the problem is easier to solve with these tests. However, we did not find a solution for the problem in the literature. De Groot [4] and Van Naerssen [11] claim to have designed a method but we did not consider their arguments convincing, in fact some are false. The method they describe is not suitable for mathematics tests. The main idea is to use many questions which nearly all the good students will answer correctly and all the poor students incorrectly and to use a number of these questions several times if on previous occasions there was a large correlation between the answers to such a question and the overall score. Both of these question types therefore occur regularly and it is suggested that the method works best if the questions remain secret. This we consider unacceptable and also impossible because the better students will remember the questions even if they are not allowed to take them out of the examination room.

We believe that it is impossible to completely avoid statistical fluctuation in the standards. A possible method is to have a very large collection of possible test-questions (known to the students) and to construct each test by taking a sample from the fixed collection. We believe that keeping the teacher 'constant' is also a fairly good guarantee for constant standards. Besides, if at the end of a year a student is judged on the basis of several examinations, the effect of fluctuations in standards of separate tests will be diminished sufficiently. An other way to decrease the effect of such fluctuations is to take the border line between passing and failing too low. The fact that (in The Netherlands) even after many years of selection a large percentage of the remaining students fails the next sequence of tests seems to indicate that this border line is indeed always taken too low.

3. A COMPARISON EXPERIMENT

Equipped with all the knowledge partially described in Section 2, we decided to compare a classical essay test (KL in the sequel) of the kind which has been used for several years with a multiple-choice test (M.C. in the sequel). It was decided that the essay test (i.e. KL) would be designed in the same way as always with as only difference that it would take 2 hours instead of 3. The subject matter was: (1) linear differential equations, (2) series.

3.1. *The Essay Test*

This test was designed by 2 members of the faculty and discussed with the different teachers of the course. After certain alterations it was given to another faculty member to judge with respect to difficulty, testing what one wishes to test, amount of time needed, etc. In a final meeting the final version of KL was agreed upon.

The test KL was as follows:

(1) (a) Solve $y''-4y'+3y=e^{3x}$.

(b) Determine all real solutions of $y''+2y'+5y=\sin 2x$.

(2) Determine for which real values of x the series

$$\sum_{n=1}^{\infty} \frac{(x^2-1)^n}{\sqrt{n}}$$

is respectively absolutely convergent, conditionally convergent, divergent.

(3) Determine whether the following series is convergent:

$$\sum_{n=1}^{\infty} \left(1-\sqrt[4]{1+\frac{1}{n}}\right).$$

(4) Calculate $\lim_{x\to 0} \dfrac{x^2e^x+\sin x \log(1-x)}{x^2 \arctan x}$.

Subsequently, we compiled the following list of facts and skills which (in our opinion) were tested by the questions of KL.

Problem

(1a) (A) setting up the characteristic equation and determining the roots,

(B) determining a special solution of the d.e. if the right-hand side is a solution of the homogeneous equation.

(1b) (C) Solving a homogeneous linear d.e. in the case where the characteristic equation has complex roots,

(D) guessing a special solution.

(2) (E) Recognizing a power series,

(F) determining the radius of convergence,

(G) checking whether an alternating series is convergent,

(H) comparison test.

(3) (I) Using the binomial series,

(H) comparison test.

(4) (J) Knowing the power series expansion of $\sin x$, e^x, $\log(1-x)$, and $\arctan x$,
(K) multiplication of power series,
(L) determining a limit using power series.

3.2. *The Multiple-Choice Test*

In the meantime several members of the faculty had tried their hand at designing m.c. questions pertaining to the subject matter of the test. It was required that each question should test one thing only and not a complex of skills. In this way a large list of possible questions had been compiled.

Using this list and if necessary (as was often the case) by designing more m.c. items we compiled a list of twelve m.c. questions testing the same skills as in the list of Section 3.1. For a number of these there were alternative questions. By the same members of the faculty as mentioned in 3.1 a choice was made from this list. The result was the following test (M.C.), (in which problem *J* also tests skill *H* and problem *K* also tests skill *J*).

A. The differential equation $y'''+3y''-4y=0$ has a characteristic equation with the roots

(1) $t_1=0$, $t_2=1$, $t_3=-4$
(2) $t_1=0$, $t_2=-1$, $t_3=4$
(3) $t_1=-1$, $t_2=2$, $t_3=-2$
(4) $t_1=1$, $t_2=-2$, $t_3=-2$.

B. Exactly one of the following eight functions is a solution of the differential equation

$$y''+2y'+y=2e^{-x}.$$

In which line is this function?

(1) $y=2e^x+3e^{-x}$; $y=(2+3x)\,e^{-x}$
(2) $y=3e^x+e^{-x}$; $y=e^{-x}(2\cos x+3\sin x)$
(3) $y=(3+x^2)\,e^{-x}$; $y=e^x(2x+3)$
(4) $y=e^{-x}+xe^{-x}$; $y=x^3e^{-x}+2e^{-x}$.

C. The complete solution of the differential equation

$$y''-2y'+2y=0$$

is

(1) $y=Ae^x\sin(x+B)$,
(2) $y=e^x(A\sin 2x+B\cos 2x)$,
(3) $y=e^{-x}(A\sin x+B\cos x)$,
(4) none of the answers (1), (2), (3) is correct.

D. To find a solution of the differential equation

$$y''+2y'+10y=\cos 3x$$

one can try one of the following functions. Which?

(1) $y=(A+Bx)e^{-x}\cos 3x$,
(2) $y=A\cos 3x+Bx\cos 3x$,
(3) $y=Ax\sin 3x+Bx\cos 3x$,
(4) $y=A\cos 3x+B\sin 3x$.

E. The series $\sum_{n=1}^{\infty}\frac{x+n}{e^{x+n}}$ (x real)

(1) is divergent for $x<0$,
(2) is convergent for all x,
(3) is convergent only if x is a negative integer,
(4) none of the answers (1), (2), (3) is correct.

F. The series $\sum_{n=1}^{\infty}\frac{x^{2n}}{2\log(n+1)}$ (x real) is convergent

(1) for $-2<x<2$,
(2) only in $x=0$,
(3) for $-1<x<1$,
(4) for $|x|<\sqrt{e}$.

G. The series $\sum_{n=1}^{\infty} n^{-1/2}\left(\frac{2x+3}{x+3}\right)^n$ is conditionally convergent for

(1) $x=-1$,
(2) $x=-2$,
(3) $x=0$,
(4) $x=1$.

H. If the series $\sum_{n=1}^{\infty} u_n$ has the property

(1) $u_n=\frac{1}{n^{1+(1/n)}}$, then the series is convergent,
(2) $\lim_{n\to\infty} n|u_n|<1$, then the series is convergent,
(3) $\lim_{n\to\infty}|nu_n|=1$, then the series is divergent,
(4) none of the answers (1), (2), (3) is correct.

I. The coefficient of x^3 in the power series expansion of $\dfrac{1}{\sqrt[3]{1-6x}}$ is

(1) $-13\frac{1}{3}$,
(2) 224,
(3) $37\frac{1}{3}$,
(4) $74\frac{2}{3}$.

J. Let $u_n = \left\{\dfrac{1}{n} + \log\left(1 - \dfrac{1}{n}\right)\right\}$. Then the series $\sum_{n=2}^{\infty} u_n$

(1) is divergent because $u_n > \dfrac{1}{n}$,

(2) is convergent because $u_n < \dfrac{1}{n^2}$,

(3) is divergent because $\lim_{n\to\infty} nu_n = 1 + e^{-1} > 0$,

(4) is convergent because $\lim_{n\to\infty} n^2 u_n$ exists.

K. The coefficient of x^5 in the power series expansion of $e^{-x^2}\arctan x$ is

(1) $\frac{11}{30}$,
(2) $\frac{31}{30}$,
(3) $\frac{49}{51}$,
(4) none of the answers (1), (2), (3) is correct.

L. $\lim_{x\to 0} \dfrac{e^x + A\cos x}{x^2}$ exists for

(1) $A=0$,
(2) $A=-1$,
(3) no value of A,
(4) all values of A.

3.3. *The Combined Test*

Normally the calculus test for the subjects treated here takes 3 hours. In this case all the students (701 students) did both parts in 4 hours. The group was split at random into 2 nearly equal parts. The first group started with M.C., the other group started with KL. After 2 hours both groups received the other half of the test after handing in their answers to the first set of questions.

In order to correct for possible effects of fatigue and to prevent harm to the students' interest caused by this experiment it was decided that for each part (i.e. KL and M.C.) the scores of the students with the lowest average would be increased in order to make the average score for the group which

made that part during the first half of the test equal to the average score for the group which had ended the test with this part. With such large groups one would safely assume that these averages would have been the same if they had done the problems at the same time.

Scoring for KL was 10 points for each of the 4 problems, total divided by 4. Scoring for M.C. was: 3 points for correct answer, −1 for incorrect answer, no answer 0 points, total divided by 3.6. The final score was obtained by first correcting for fatigue as described above, then taking the average of the 2 parts, rounded upwards.

Grading for KL was done in the standard manner of our department, i.e. a grading system was set up beforehand and each paper was read by two persons. As usual the correlation coefficient for these readings was very high (≈ 0.95).

4. Results and conclusions

4.1. *Fatigue Correction*

Group 1 (size 364) starting with M.C. scored an average $m=5.73$ (standard deviation $s=1.9$). Group 2 (size 337) ending with M.C. scored $m=5.71$ (standard deviation $s=1.9$). The same numbers for KL are

group 1; $m=6.32$ ($s=1.9$)
group 2; $m=6.68$ ($s=2.0$).

TABLE I

Scoring for M.C.				Scoring for KL			
Score	Group 1	Group 2	Group 1+2	Group 1	Group 2	Group 1+2	Score
0	3	1	4	2	0	2	0
1	10	8	18	0	2	2	1
2	6	15	21	7	6	13	2
3	27	19	46	18	14	32	3
4	58	50	108	31	25	56	4
5	22	31	53	44	30	74	5
6	79	68	147	80	56	136	6
7	85	81	166	73	81	154	7
8	53	43	96	64	58	122	8
9	16	17	33	27	37	64	9
10	5	4	9	18	28	46	10
Total	364	337	701	364	337	701	Total
m	5.73	5.71	5.72	6.32	6.68	6.49	mean
s	1.92	1.90	1.91	1.88	1.95	1.92	st. dev.

Although the computed deviations show that one can hardly see any effect of fatigue the students in group 1 all received 0.36 extra points for KL before the final score was computed (see Table I).

4.2. *Comments on M.C.*

The correlation between separate questions and the total score for M.C. was computed. We found that question A contributes very little to the total score. Clearly the question was too simple. Questions C, H, and J were apparently hard (see Table II). Most of the computed correlations are high, i.e. the questions were good questions. They are of the kind discussed in Section 2.5.

TABLE II

Average score for the separate problems

Problem	Group 1	Group 2	Total
KL 1A	4.62	4.67	4.64
KL 1B	3.97	4.00	3.98
2	6.38	6.53	6.45
3	2.60	3.30	2.94
4	7.69	8.21	7.94
M.C. A	9.27	9.22	9.24
B	6.57	6.20	6.39
C	−0.85	−1.26	−1.05
D	5.83	6.77	6.29
E	8.05	8.49	8.26
F	8.67	8.42	8.55
G	7.93	8.06	7.99
H	1.36	1.08	1.23
I	7.85	6.89	7.39
J	−0.43	−0.27	−0.35
K	6.18	6.35	6.26
L	8.30	8.58	8.43

4.3. *Correlation*

Our original aim was to compare corresponding parts of the two tests and the total results of the tests. The correlation coefficients ρ which we found are given below:

M.C.	KL	ρ
A, B	1a	0.18
C, D	1b	0.26
E, F, G	2	0.29
H, I, J	3	0.21
J, K, L	4	0.40
total scores		0.57

It turned out that the numbers ρ for similar questions were not significantly larger than those for non-corresponding questions in the two parts.

Is the number 0.57 large enough to say that M.C. is a good test, i.e. a test measuring those skills which we wish to measure? If a good test is repeated and there are no memory effects one expects something like a correlation coefficient of 0.70. Since a test should have a lot more than the 12 questions which we had, a somewhat smaller number is reasonable. Therefore it is our opinion that the 2 tests, as far as measuring the effect of the educational program in differential equations and series, are not significantly different. Our conclusions are

(1) A large part of a calculus course can be tested quite well by a M.C. test.

(2) An essay test is easier to design and works just as well (for all of the parts of the course).

(3) A mixture of essay test and short answer test is probably the best method.

(4) Constant standards can be obtained only by a *constant* very large set of questions. (This is an opinion the authors have formed as a result of their research.)

(5) The time saved by machine correction of an M.C. test is tremendous.

(6) It seems likely that very few teachers are able te design suitable multiple choice questions (and especially suitable answers). The time lost in designing a test like we did is tremendous.

In our opinion the net effect of (5) and (6) is a saving of time but a big increase of the probability that the test is poor.

T.H., Eindhoven

BIBLIOGRAPHY

[1] Ahmann, J. S., 'Testing Student Achievements and Aptitudes', The Center for Applied Research and Education, Inc., Washington 1962.

[2] Bloom, B. S. (ed.), *Taxonomy of Educational Objectives, I,* Longman, Green and Co., New York 1956.

[3] Ebel, R. L. and Dora, E. Damrin, 'Tests and Examinations', chapter from *Encyclopedia of Educational Research,* Mcmillan, New York 1960.

[4] Groot, A. D. de, *Vijven en Zessen,* Wolters, Groningen 1966.

[5] Groot, A. D. de, A lecture held at the Technological University Eindhoven.

[6] Husén, T. (ed.), *International Study of Achievement in Mathematics,* John Wiley & Sons, New York, 1967.

[7] Kamps, H. J. L. and J. H. van Lint, *Objectieve toetsen?* T.H. Report, 69-WSK-05, Technological University Eindhoven.

[8] Klerk, L. F. W. de, *Objectieve Studietoetsen,* Stichting VAM, Voorschoten.

[9] Lang, G., *Inleiding over Studietoetsen,* Stichting Research Instituut voor de Toegepaste Psychologie of the University of Amsterdam.

[10] Lindquist, E. F. (ed.), *Educational Measurement,* American Council on Education, Washington 1951.
[11] Naerssen, R. F. van, 'Het handhaven van eenmaal aangenomen normen bij opeenvolgende objectieve toetsen', *Paedagogische Studiën* **43** (1966), 312–320.
[12] Nunnally, J. C., *Educational Measurement and Evaluation,* McGraw-Hill, New York 1964.
[13] Wood, R., 'Objectives in the Teaching of Mathematics', *Educational Research* **10** (1968), 83–98.